HIERARCHICAL FORECASTING FOR PREDICTING SPARE PARTS DEMAND IN THE SOUTH KOREAN NAVY

A thesis submitted for the degree of

Doctor of Philosophy

by

Seongmin Moon

Newcastle University Business School

July 2010

Abstract

In the South Korean Navy the demand for many spare parts is infrequent and the volume of items required is irregular. This pattern, known as non-normal demand, makes forecasting difficult. This research uses data obtained from the South Korean Navy to compare the performance of forecasting methods that use hierarchical and direct forecasting strategies for predicting the demand for spare parts.

Among various forecasting methods tested, a simple combination of exponential smoothing models, which uses a hierarchical forecasting strategy, was found to minimise forecasting errors and inventory costs. This simple combination forecasting method was generated by a simple combination between an exponential smoothing model with quarterly aggregated data adjusted for linear trend at group level and an exponential smoothing model with monthly aggregated unadjusted data at item level.

Logistic regression classification model for predicting the relative performance of alternative forecasting methods (i.e. a direct forecasting method vs. a hierarchical forecasting method) by multivariate demand features of spare parts was developed. Logistic regression classification model is generalisable, because it is based on relationships between the relative performance of alternative forecasting methods and demand features. This classification model reduced inventory costs, compared to the results of only using the simple combination forecasting method. This classification model is likely to be a promising model to guide the selection of a forecasting method between alternative forecasting methods for predicting spare parts demand in the South Korean Navy, so that it could maximise the operational availability of weapon systems.

Acknowledgements

I am grateful to the Republic of Korean Navy for sponsoring me to study a Ph.D. programme. I wish to express my sincere gratitude to my supervisors, Dr. Andrew Simpson and Prof. Christian Hicks, for their trust, guidance and full support throughout the period of this research. I would like to acknowledge to my colleagues in the Navy and Newcastle University Business School. I have to give a special thank you to my family and Warrin Laoponghanr for support and inspiration.

Table of Contents

Abstract i

Acknowledgements ii

Table of Contents iii

List of Tables ix

List of Figures xiv

Glossary of Abbreviations xvi

Chapter 1. Introduction 1

1.1 Background 1

1.1.1 Operational availability 1

1.1.2 Budgetary problem 3

1.2 Research Problem 5

1.2.1 Forecasting accuracy for the spare parts demand in the South Korean Navy 7

1.2.2 Nature of spare parts demand 8

1.2.3 Hierarchical structure 11

1.2.4 Forecasting non-normal demand 12

1.3 Aim and Objectives 14

1.4 Research Questions 15

1.5 Research Gaps and Contributions 16

1.6 Outline of the Thesis 19

Chapter 2. Literature Review 21

2.1 Forecasting Methods in General 21

2.2 Direct Forecasting 23

2.2.1 Transformation for forecasting *23*

2.2.2 Multivariate versus univariate forecasting methods *26*

2.2.3 Direct forecasting for non-normal demand *27*

2.3 Hierarchical Forecasting 40

2.3.1 Concept of hierarchical forecasting *40*

2.3.2 Bottom-up and top-down *41*

2.3.3 Proration methods *42*

2.4 Performance of Forecasting 48

2.4.1 Bottom-up forecasting *52*

2.4.2 Top-down forecasting *53*

2.4.3 Combinatorial forecasting *56*

2.4.4 Disagreement in the performance *58*

2.4.5 Hierarchical forecasting for non-normal demand *58*

2.5 Classification of Demand for Forecasting 60

2.5.1 Direct forecasting *61*

2.5.2 Hierarchical forecasting *70*

2.5.3 Evaluation of hierarchical forecasting research *80*

2.6 Measures of Forecasting Accuracy 85

2.6.1 Absolute measures of accuracy *86*

2.6.2 Relative measures of accuracy *89*

2.6.3 Derivative measures of accuracy *92*

2.7 Derivative Measures of Inventory Model Performance 94

2.7.1 Inventory classification *95*

2.7.2 Inventory systems *96*

2.7.3 Measures *101*

2.8 Summary and Conclusion 104

Chapter 3. Methodology 110

3.1 Purpose of Research 110

3.2 Empirical Research in Operations Management 111

3.3 Case Study Strategy 113

3.4 Theory in Case Study 114

3.4.1 Theory as knowledge base 115

3.4.2 Theory induction from data 116

3.4.3 Generalisation 116

3.5 Research Design 117

3.5.1 Four components of research design 118

3.5.2 Single case design 120

3.6 Evaluation of Research 122

3.6.1 Construct validity 122

3.6.2 Internal validity 123

3.6.3 External validity 124

3.6.4 Reliability 125

3.7 Research Procedure 125

3.8 Summary and Conclusion 127

Chapter 4. Nature Of The Spare Parts Demand 129

4.1 General Information 130

4.2 Inventory Control Method 132

4.2.1 Authorised Stock Item List 132

4.2.2 Inventory objective of Authorised Stock Item List items 133

4.3 Spare Parts Demand Analysis 135

4.3.1 National Stock Number 136

4.3.2 Data selection and grouping *137*

4.3.3 Measures of demand features *141*

4.3.4 Historical demand analysis *146*

4.3.5 Seasonality *159*

4.3.6 Procurement lead time *162*

4.3.7 Item unit price *164*

4.4 Decomposition Procedure 167

4.5 Sources of Non-Normality 174

4.5.1 A few large customers *174*

4.5.2 Operating factors *175*

4.5.3 Multi-echelon inventory systems *176*

4.5.4 Budgeting process *179*

4.5.5 Maintenance system *180*

4.6 Summary and Conclusion 182

Chapter 5. Forecasting The Spare Parts Demand 185

5.1 Forecasting Spare Parts Demand in the South Korean Navy 186

5.2 Absolute and Relative Measures of Accuracy 189

5.3 Derivative Measures of Accuracy 192

5.3.1 Simulation model structure *193*

5.3.2 Simulation process *196*

5.3.3 Measures of inventory model performance *199*

5.3.4 Simulation model selection *203*

5.4 Direct Forecasting 205

5.4.1 The development of direct forecasting methods *206*

5.4.2 Group level forecasting *213*

5.4.3 Item level forecasting 217

5.4.4 Simulation 223

5.5 Hierarchical Forecasting 227

5.5.1 The development of hierarchical forecasting methods 228

5.5.2 Forecasting performance comparisons 230

5.5.3 Simulation 241

5.6 Summary and Conclusion 255

5.6.1 Summary of findings 255

5.6.2 Forecasting scheme for the South Korean Navy 257

5.6.3 Conclusion 258

Chapter 6. Forecasting Performance And Demand Features 260

6.1 Competing Performance 260

6.2 Possible Demand Features to Guide the Selection of a Forecasting Method 261

6.3 Process of Classification 266

6.3.1 Forecasting performance 266

6.3.2 Various classification methods 269

6.3.3 Classification by logistic regression 271

6.3.4 Diagnostics 275

6.3.5 Cross-validation 277

6.3.6 Final model 287

6.4 Classification Results 289

6.4.1 Classification by the logistic regression model 289

6.4.2 Interpretation of the classification model 292

6.4.3 Predictors in the classification model 294

6.4.4 Individual relations between forecasting methods and demand features 298

6.5 Summary and Conclusion 306

6.5.1 Summary of findings *306*

6.5.2 Forecasting scheme for the South Korean Navy *310*

6.5.3 Conclusion *311*

Chapter 7. Conclusion 312

7.1 Nature of the Spare Parts Demand 312

7.2 The Performance of the Forecasting Methods 314

7.3 Forecasting Performance and Demand Features 316

7.4 Contributions 318

7.5 Forecasting Scheme for the South Korean Navy 321

7.6 Limitations 322

7.7 Further Research 323

Appendices 325

Appendix A - Probability Distributions 325

Appendix B - NATO Supply Group 328

Appendix C - Time Series Plots Of The Sums Of The Monthly Aggregated Time Series Of The Spare Parts Demand For The Eight Pieces Of Equipment 330

Appendix D - Hierarchical Forecasting Comparisons (MAD) 334

Appendix E - Hierarchical Forecasting Comparisons (RMSE) 338

Appendix F - Hierarchical Forecasting Comparisons (2005 ~ 2007) 343

Appendix G - Hierarchical Forecasting Simulation Results Comparisons (2005 ~ 2007) 348

References 353

List of Tables

Table 1-1 Forecasting accuracy in terms of item 7

Table 1-2 Forecasting accuracy in terms of volume 8

Table 1-3 Classification of spare parts in the South Korean Navy 9

Table 2-1 A review of direct forecasting methods for non-normal demand 31

Table 2-2 A review of forecasting strategies and forecasting performance 50

Table 2-3 Classification of spare parts demand by two dimensions 62

Table 2-4 Classification of demand 67

Table 2-5 Classification scheme for identifying intermittence 69

Table 2-6 The influence of demand features upon the relative performance of top-down forecasting and direct forecasting 71

Table 2-7 Research on the influence of correlations upon hierarchical forecasting 81

Table 2-8 Weapon System Indicator Code 95

Table 2-9 The performance of inventory systems for spare parts 100

Table 3-1 Research category in operations management 113

Table 3-2 The research questions 114

Table 3-3 Choice of number of cases 120

Table 3-4 Five rationales for single case design 121

Table 4-1 Classification of spare parts 131

Table 4-2 Types of Demand 133

Table 4-3 Inventory objective of the South Korean Navy 134

Table 4-4 Selection of data (300 items) 138

Table 4-5 Average statistics for the monthly aggregated data 153

Table 4-6 Statistical features of the time series 157

Table 4-7 Test for quarterly seasonal effect 160

Table 4-8 Quarterly seasonal effect in equipment groups 161

Table 4-9 Test for monthly seasonal effect 161

Table 4-10 Monthly seasonal effect in equipment groups 162

Table 4-11 PROLT analysis 163

Table 4-12 Item unit price and DV 165

Table 4-13 Inconsistency between the number of items ordered in the South Korean Navy 179

Table 5-1 Assumptions for the simulation 195

Table 5-2 Example of the simulation process 197

Table 5-3 Simulation model comparisons 203

Table 5-4 Sums of direct forecasting errors at group level 214

Table 5-5 Mean ranks of direct forecasting methods at group level 216

Table 5-6 Forecasting accuracy comparisons by data aggregation (group) 216

Table 5-7 Forecasting accuracy comparisons by data adjustment (group) 217

Table 5-8 Sums of forecasting errors (item) 218

Table 5-9 Mean ranks of direct forecasting methods at item level 219

Table 5-10 Forecasting accuracy comparisons by aggregation (item) 219

Table 5-11 Forecasting accuracy comparisons by data adjustment (item) 220

Table 5-12 Mean ranks of direct forecasting methods for equipment groups at group level 221

Table 5-13 Mean ranks of direct forecasting methods for equipment groups at item level 222

Table 5-14 Simulation results for direct forecasting 224

Table 5-15 Inventory costs and mean ranks of direct forecasting 224

Table 5-16 Simulation results of direct forecasting for the equipment groups 225

Table 5-17 Mean ranks of direct forecasting methods for equipment group in terms of inventory costs and MAD 226
Table 5-18 Robust direct forecasting methods 226
Table 5-19 Top 21 hierarchical forecasting methods (MAD) 231
Table 5-20 Top 21 hierarchical forecasting methods (RMSE) 232
Table 5-21 Top 21 hierarchical forecasting methods (mean rank) 234
Table 5-22 Direct forecasting methods used for the top 21 hierarchical forecasting methods 235
Table 5-23 Proration methods comparisons [LN(HF/*tsm*)] 236
Table 5-24 Proration methods comparisons in the 3 equipment groups [LN(ratio)] 237
Table 5-25 Top 10 hierarchical forecasting methods for Gun/RD in terms of either the natural log relative error for MAD or RMSE 239
Table 5-26 Top 10 hierarchical forecasting methods for ME in terms of either the natural log relative error for MAD or RMSE 239
Table 5-27 Top 10 hierarchical forecasting methods for GE/AC in terms of either the natural log relative error for MAD or RMSE 240
Table 5-28 The performance of hierarchical forecasting methods in terms of absolute and relative measures 241
Table 5-29 Top 20 hierarchical forecasting methods 242
Table 5-30 Direct forecasting methods used for the top 21 or the top 20 hierarchical forecasting methods 244
Table 5-31 The 33 hierarchical forecasting methods in simulation 246
Table 5-32 Mean rank of the 33 hierarchical forecasting methods 247
Table 5-33 Proration methods comparisons (simulation) 248
Table 5-34 Proration methods comparisons in the 3 equipment groups (simulation) .. 249

Table 5-35 Top 10 hierarchical forecasting methods for Gun/RD in terms of either the natural log relative error for MAD or RMSE, or the total inventory costs 251

Table 5-36 Top 10 hierarchical forecasting methods for ME in terms of either the natural log relative error for MAD or RMSE, or the total inventory costs 252

Table 5-37 Top 10 hierarchical forecasting methods for GE/AC in terms of either the natural log relative error for MAD or RMSE, or the total inventory costs 253

Table 5-38 The performance of hierarchical forecasting methods 254

Table 5-39 Summary of the performance of forecasting methods............................... 256

Table 6-1 Possible variables to guide the selection of a forecasting method 263

Table 6-2 Forecasting performance comparisons in terms of the sum of absolute deviations divided by the monthly mean.. 268

Table 6-3 Categorical variables coding 272

Table 6-4 Predictors contained within each cross-validation set................................ 278

Table 6-5 The performance of backward and forward stepwise methods 280

Table 6-6 Bias comparisons between forward and backward stepwise........................ 281

Table 6-7 Performance comparisons among forward, backward and *SCtqum* 282

Table 6-8 Bias comparisons among forward, backward and employed 283

Table 6-9 Classification results using the employed model 284

Table 6-10 Prediction results 284

Table 6-11 Multicollinearity statistics... 288

Table 6-12 Prediction results from resubstitution.. 289

Table 6-13 Variables in the equation... 292

Table 6-14 Relations between predictors and odds of classification in the models 295

Table 6-15 Cross-tabulation between the alternative forecasting methods and the three equipment groups 299

Table 6-16 Chi-Square test between the forecasts and the equipment groups.............. 300

Table 6-17 ANOVA for the variables in the logistic regression model......................... 301

Table 6-18 Description of I.Cv(size) 303

Table 6-19 Description of I.Kurtosis 303

Table 6-20 Correlation between I.Cv(size) and I.Kurtosis 303

Table 6-21 Individual relations between forecasting methods and variables 306

Table 6-22 The influence of demand features upon the performance of the alternative forecasting methods... 309

Table 7-1 A summary of findings: the performance of forecasting methods................ 315

Table 7-2 A summary of findings: the influence of demand features upon the relative performance of alternative forecasting methods 318

Table 7-3 Contributions of the research... 320

Table A-1 Probability distributions.. 326

Table B-2 NATO Supply Group 328

Table D-3 Hierarchical forecasting comparisons in term of MAD 334

Table E-4 Hierarchical forecasting comparisons in term of RMSE 338

Table F-5 Hierarchical forecasting comparisons in term of the mean rank of MAD and RMSE 343

Table G-6 Hierarchical forecasting simulation results 348

List of Figures

Figure 1-1 Trade-off between reliability and maintainability 3

Figure 1-2 The US Department of Defence budget profile 4

Figure 2-1 Two-dimensional classification of demand during lead time 66

Figure 3-1 Research procedure 126

Figure 4-1 Mean annual demand for items in three value groups 132

Figure 4-2 National Stock Number 136

Figure 4-3 Grouping structure 139

Figure 4-4 Time series plot of the sum of the 300 monthly aggregated spare parts demand time series 147

Figure 4-5 Histogram of the sum of the 300 monthly aggregated spare parts demand data 148

Figure 4-6 Probability plot of the sum of the 300 monthly aggregated spare parts demand data 148

Figure 4-7 Time series plots of the sums of the monthly aggregated time series of the spare parts demand for the eight pieces of equipment 150

Figure 4-8 Histograms of the sums of the monthly aggregated time series of the spare parts demand for the eight pieces of equipment 151

Figure 4-9 Probability plots of the sums of the monthly aggregated time series of the spare parts demand for the eight pieces of equipment 152

Figure 4-10 Histograms of PROLT 163

Figure 4-11 Histograms of item unit price 165

Figure 4-12 Histograms of historical dollar volume 166

Figure 4-13 Yearly data adjustment 171

Figure 4-14 Quarterly data adjustment 172

Figure 4-15 Monthly data adjustment 172

Figure 4-16 Ten direct forecasting methods 173

Figure 4-17 Abbreviation for a hierarchical forecasting method 173

Figure 4-18 Change of demand (N-ASL → ASL) 177

Figure 4-19 Change of demand (N-ASL → ASL → N-ASL) 178

Figure 5-1 Simulation process 194

Figure 5-2 Forecasting performance comparisons in the two simulation models 204

Figure 5-3 Residuals against fitted values 207

Figure 5-4 Standardized residuals 208

Figure 5-5 Probability plots of the residuals 209

Figure 6-1 Relative forecasting performance (*tsm* vs. *SCtqum*) 261

Figure 6-2 Abbreviation for a variable 262

Figure 6-3 Forecasting performance comparisons 269

Figure 6-4 Forecasting performance improvement of the employed model 286

Figure 6-5 Forecasting performance improvement in the model 291

Figure 6-6 Relations between *SCtqum-tsm* and I.Cv(size) 304

Figure 6-7 Relations between *SCtqum-tsm* and I.Kurtosis 304

Figure C-1 Time series plot of spare parts for Gun I 330

Figure C-2 Time series plot of spare parts for Gun II 330

Figure C-3 Time series plot of spare parts for Gun III 331

Figure C-4 Time series plot of spare parts for Main Engine I 331

Figure C-5 Time series plot of spare parts for Main Engine II 332

Figure C-6 Time series plot of spare parts for Generator I 332

Figure C-7 Time series plot of spare parts for Air Compressor I 333

Figure C-8 Time series plot of spare parts for Radar I 333

Glossary of Abbreviations

Forecasting method

AR	AutoRegressive
ARIMA	AutoRegressive Integrated Moving Average
ARMA	AutoRegressive Moving Average
BU	Bottom-up Forecasting
CF	Combinatorial Forecasting
DF	Direct Forecasting
DRF	Derived Forecasting
ES	Exponential Smoothing
EWMA	Exponentially Weighted Moving Average
HF	Hierarchical Forecasting
MA	Moving Average
SBA	Syntetos-Boylan Approximation
SC	Simple Combination
SES	Simple Exponential Smoothing
SMA	Simple Moving Average
TD	Top-down Forecasting
TD1	Top-down Forecasting1
TD2	Top-down Forecasting2
VARMA	Vector Mixed AutoRegressive/ Moving Average
WC	Weighted Combination
WMA	Weighted Moving Average

Decomposition

u	A forecast with unadjusted data
t	A forecast with data adjusted for linear trend
s	A forecast with data adjusted for additive seasonality
ts	A forecast with data adjusted for linear trend and additive seasonality
m	A forecast with monthly aggregated data
q	A forecast with quarterly aggregated data
y	A forecast with yearly aggregated data

Demand feature

ADI	Average inter-Demand Interval
Corr(group)	Correlations of the item level time series with the group level time series
Corr(item)	Correlations of the item level time series with other item level time series in the same group
Cv(size)	Coefficient of variation in demand size
Pr(peak)	Proportion of peak demands
Pr(zero)	Proportion of zero demand periods
PROLT	Procurement Lead Time
DV	Historical Dollar Volume
SI	Item Demand Series' Seasonal Index
UV	Historical Unit Volume
HP	Historical Item Forecasting Performance
G (prefix)	Quarterly aggregated data at group level
I (prefix)	Monthly aggregated data at item level

Error measure

GRMSE	Geometric Root Mean Squared Error
LN(ratio)	Natural log relative error [$\ln(\text{error}_{HF}/\text{error}_{DF})$]
MAD	Mean Absolute Deviation
MAD/A	Mean Absolute Deviation Average
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error
RC	Average Inventory Cost Regret
RS	Average Service level Regret
RGRMSE	Relative Geometric Root Mean Squared Error
RMSE	Root Mean Squared Error
RMSEP	Prorated Forecast RMSE
RSFE	Running Sum of Forecast Error
S	Relative Size of RSFE (RSFE/MAD)
SSE	Sum of Squared Error

Miscellanea

AC	Air Compressor
AD	Annual Demand
ANOVA	Analysis Of Variance
ASL	Authorised Stock Item List
GE	Generator
IIN	Item Identification Number
ME	Main Engine
MND	The Korean Ministry of National Defence
MTBF	Mean Time Between Failure
MTBM	Mean Time Between Maintenance
MTBO	Mean Time Between Overhaul
MTBR	Mean Time Between Repair
N-ASL	Non-Authorised Stock Item List
NATO	North Atlantic Treaty Organisation
NBC	Nuclear Biological Chemical weapon
NCB	National Codification Bureau
NIIN	National Item Identification Number
NLC	The Naval Logistics Command
NSC	NATO Supply Class
NSCG	NATO Supply Classification Group
NSG	NATO Supply Group
NSN	National Stock Number
OL	Operating Level
OST	Order & Shipping Time
PC	Procurement Cycle Quantity
RD	Radar
RO1	Requirement Objective
RO2	Requisition Objective
Seffect	Seasonal Effect
SL	Safety Level
Std	Standard Deviation
TOD	Types Of Demand

Chapter 1. Introduction

This chapter begins by describing the background of the research problems. In Section 1.2, the research problems are identified. In Section 1.3, the aim and the objectives are presented. In Section 1.4, the research questions are presented. In Section 1.5, the research gaps and the contributions are presented. Finally, this thesis is summarised in Section 1.6.

1.1 Background

After the end of the Cold War, the chances of a full-scale war happening were reduced. However, transnational or non-military threats such as terrorism and Nuclear Biological Chemical (NBC) weapon proliferation are continuously afflicting the peace of the world (Sipila, 2004). In addition, the costs of maintenance for weapon systems are increasing continuously with the development of technology and the increased complexity of weapon systems (Choi et al., 2005). The South Korean Navy is required to manage its supply system effectively in order to maximise the operational availability of warships under such conditions.

> One of the basic courses of defence policy of the Korean Ministry of National Defence (MND) is to transform its management system to a highly efficient management system with high cost effectiveness within the limited defence resources (Korean Defence White Paper, 2008, p.79).

1.1.1 Operational availability

Operational availability is defined as "the total time during the mission's duration when the weapon system or equipment is capable of meeting specified performance standards, that is, the ratio of time available when needed to total time needed" (The US Naval Institute, 1977, p. 260). A weapon system is defined as "a combination of one or more weapons with all related equipment, materials, services, personnel, and means of delivery and deployment (if applicable) required for self-sufficiency" (Schmitt, 2005, p. 456). It can refer to an individual platform such as a warship, a submarine or an aircraft (The US Naval Institute, 1977). Although the South Korean Navy has classified regulations for weapon systems that specify the required operational availability, the operational availability of the Naval warships was reported to be an average of 80% in 2005 (Lee, 2007).

The operational availability of weapon systems depends on reliability, maintainability and supply (The US Naval Institute, 1977). Reliability refers to the probability that a system will perform as expected during the entire period of a mission, and can be gauged in terms of mean time between failure (MTBF) (The US Naval Institute, 1977). Maintainability refers to the probability that a system, inoperable for any reason, can be returned to service in a given period (The US Naval Institute, 1977). The US Naval Institute (1977) argued that maintainability, reliability and supply are important factors that determine operational availability. It was postulated that, if operational availability is quantified with the spare parts supply fixed for a weapon system, both higher reliability and higher maintainability for the weapon system cannot be achieved at the same time, because the reliability may have to be traded off in view of the maintainability collations. The higher reliability of a weapon system requires a lower maintainability of the weapon system (i.e. a longer time to return the system to service). Figure 1-1 illustrates that system I has longer MTBF (i.e. higher reliability) but requires a much longer repair time to return the system to service (i.e. lower maintainability) than system II. In order to achieve both higher reliability and higher maintainability (so as to maximise the operational availability of weapon systems), an adequate supply of spare parts to meet the requirements of repair and maintenance is necessary (The US Naval Institute, 1977).

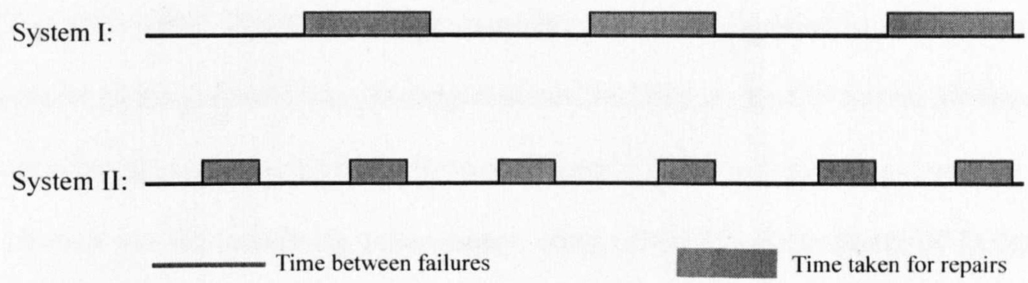


Figure 1-1 Trade-off between reliability and maintainability
(The US Naval Institute, 1977)

The operational availability of weapon systems is important and should not be jeopardised by lack of spare parts, because the operational availability can affect the operations of militaries which are charged with the defence of the nation (Fowler, 2003). Thus, it is common for military forces to hold a large stock of spare parts. For example, the US Department of Defence continues to hold a 60% excess of spare parts inventories which are expected to be required at any present time (Hinton Jr., 1999). The British Navy holds almost twice as many spare parts as are expected to be required (Fowler, 2003). However, militaries could not hold unlimited amount of spare parts for sustaining operational availability. This is due to budgetary limitations. These budgetary problems are related as follows.

1.1.2 Budgetary problem

Militaries in some countries are suffering from limited budgets to maintain operational availability (Rustenburg et al., 2001, Agripino et al., 2002, Yonhap News, 2007, Lee, 2007). For instance, the US Department of Defence faces escalating maintenance costs (Agripino et al., 2002). Figure 1-2 shows that the projected percentage of operational maintenance costs, within the total budget, increases continuously and will equal the total current budget (net present value) of the US Department of Defence (DoD) by the year 2024. This projection is based upon a monetary unit - billion dollars in the Fiscal Year (FY) 1999. Whilst the budget (supply) of the US Department of Defence keeps to current planned levels, the funding required to support the US forces (demand) based upon the Quadrennial Defence Review Report (QDR) is expected to increase steadily. The reasons for increasing maintenance costs of the US Department of Defence were analysed and categorised according to: a) increased operational tempo; b) increased operational requirements; c) increased life extension of existing weapon systems due to delays in new system acquisition; d) unforeseen support problems associated with aging weapon systems; and e) material shortages because of diminishing manufacturing resources and technological obsolescence (Agripino et al., 2002).

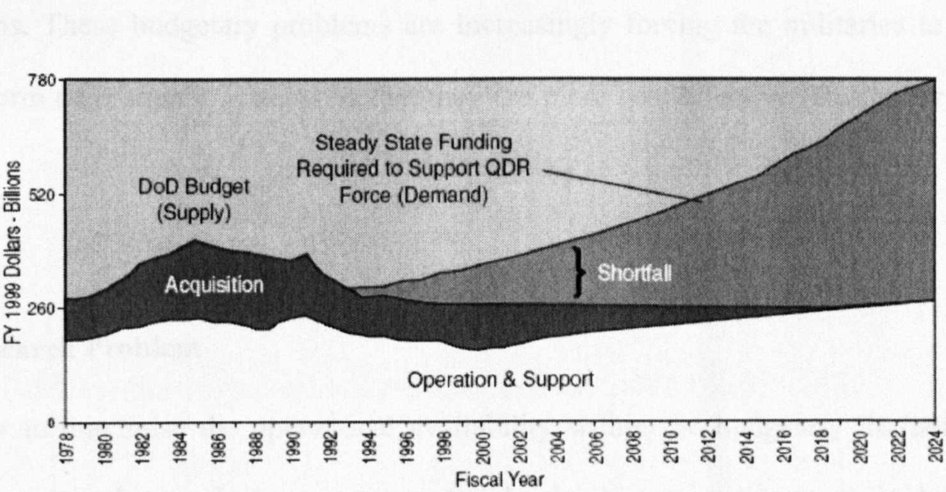


Figure 1-2 The US Department of Defence budget profile (Agripino et al., 2002)

The South Korean Air Force was reported to have suffered from the insufficient budget to sustain operational availability at a regulated level (Yonhap News, 2007). The expenditure for the maintenance of military equipment has been increased by an annual average of 0.5% from 2000 to 2006, whilst the total defence budget has been increased by an annual average of 7.3% during the same period. Although the number of items of spare parts stocked by the South Korean Air Force has increased by 53% (from 174,848 to 268,522), the budget for the spare parts has increased by only 0.7% since 2000 (Yonhap News, 2007). As a result of an insufficient budget, the operational availability of the South Korean Air Force decreased from 89.3% in 2000 to 77.8% by the first quarter of the year 2006 (Yonhap News, 2007).

The Netherlands Navy has also received an insufficient re-supply budget for spare parts to satisfy requests from the Naval warships (Rustenburg et al., 2001). It seems likely that the budgetary problem of maintaining operational availability might not be a problem only for the Netherlands Navy, the South Korean Air Force and the US Department of Defence. Most of militaries are probably suffering from budgetary problems. These budgetary problems are increasingly forcing the militaries to review and reform their supply systems so that they are more cost effective (Rustenburg et al., 2001).

1.2 Research Problem

In order to maximise the operational availability within the budgetary limitation, the spare parts supply may have to increase. Stock, also known as inventory, is defined as "the stored accumulation of transformed resources in a process" (Slack et al., 2004, p. 774). Over-stocking is usually not the best solution to the supply problem; spare parts can become obsolete or damaged; some items can deteriorate or have limited shelf life and there is also the possibility of theft or loss. Over-stocking increases inventory, the amount of capital required and carrying costs (Waller, 2003, Slack et al., 2004). Understocking can lead to a weapon system being unavailable when a timely supply of spare parts for repair and maintenance for the weapon system is obstructed by a lack of spare parts. This unavailability of the weapon system could be a waste of budget resources and could even lead to a military defeat that could cause casualties and deaths (MacDonald, 1997). A weapon system undergoing service or maintenance could remain unavailable until the spare parts are supplied (Rustenburg et al., 2000). The situation of unavailability could continue longer with a long procurement lead time.

As stated above, it is common for military forces to hold a large stock of spare parts, with often little or no demand for a large proportion of the stock items. However, the shortage of some items of spare parts is unavoidable. For example, the US Department of Defence holds a 60% excess of spare parts, with 18% of the inventory ($1.5 billion) having no demand. However, inventory shortages for some items still occur (Hinton Jr., 1999).

Forecasting is defined as "the prediction of values of a variable based on known past values of that variable or other related variables" (Hyndman et al., 1998, p. 599). Forecasting is important for supply, because forecasting is the most inexact function in supply as well as the trigger which sets the supply in motion (Waller, 2003). An accurate demand forecasting is required in order to purchase the exact amount of spare parts for the requirement of a weapon system.

Hinton (1999) claimed that the problems relating to spare parts inventory in the US Department of Defence arose because of the inaccurate forecasting of inventory requirements. Although the procurement of spare parts may be initiated to meet specific requests, it is common for requirements to change after items have been ordered (Hinton Jr., 1999). It has also been reported that South Korean military experiences spare parts supply problem caused by inaccurate forecasts of spare parts demand (Lee, 2007, Yoon and Sohn, 2007, Seon and U, 2009). This research focuses on the forecasting problems, within the context of the spare parts demand for the South Korean Navy.

1.2.1 Forecasting accuracy for the spare parts demand in the South Korean Navy

The forecasting accuracy for the spare parts demand in the South Korean Navy has not been satisfactory. Table 1-1 presents the forecasting accuracy for the spare parts demand in terms of item in the South Korean Navy.

Table 1-1 Forecasting accuracy in terms of item (Seon and U, 2009)

Forecast	Observed	2006	2007	2008
$\hat{y}_i = 0$	$y_i = 0$	**31,447**	**27,333**	**35,043**
$\hat{y}_i = 0$	$y_i \geq 1$	6,040	13,913	15,278
$\hat{y}_i \geq 1$	$y_i = 0$	13,795	3,850	4,617
$\hat{y}_i \geq 1$	$y_i \geq 1$	**10,610**	**11,421**	**11,137**
Forecasting accuracy		68%	69%	70%

Key: y_i (or $\hat{y}_i$) = the observed (or forecast) demand for item i; the numbers of correct cases are shown in bold.

In 2008, 15,278 items of spare parts, which had been forecast to be non-demanded, were demanded; 4,617 items of spare parts, which had been forecast to be demanded, were not demanded. The forecasting accuracy was calculated as "the number of correct cases divided by the number of all cases". For example, in 2008 the forecasting accuracy was calculated as 46,180 (no. of correct cases) divided by 66,075 (no. of all cases). If the forecasting accuracy is calculated in terms of volume, the forecasting accuracy is very low. Table 1-2 presents the forecasting accuracy for spare parts based on the cases with

$\hat{y}_i \geq 1$ (i.e. the cases in the second and third rows from the bottom in Table 1-1) in terms of volume in the South Korean Navy (Seon and U, 2009). Very low percentages (6 ~ 9%) of cases were correct.

Table 1-2 Forecasting accuracy in terms of volume (Seon and U, 2009)

		2006	2007	2008
Correct	$y_i = \hat{y}_i$	6%	7%	9%
Over-forecast	$y_i < \hat{y}_i$	31%	30%	30%
Under-forecast	$y_i > \hat{y}_i$	27%	38%	32%
Non-demand	$y_i = 0$	36%	25%	29%

Key: y_i (or $\hat{y}_i$) = the observed (or forecast) demand for item i; $\hat{y}_i \geq 1$.

Some authors have claimed that the reason for the inaccurate forecasts of spare parts demand in the South Korean military is that it uses simple forecasting methods which could not reflect the characteristics of spare parts demand (Choi et al., 2005, Seon and U, 2009). The identification of the characteristics of spare parts demand might be important for the development of an accurate forecasting method.

1.2.2 Nature of spare parts demand

There are two types of military spare parts, consumable and repairable (Rustenburg et al., 2001). Repairable spare parts remain in the inventory list until they are repaired and reissued or until it is decided to remove them. Consumable spare parts are deleted from the inventory list when they are supplied to users. Spare parts are classified into parts, components and assemblies in the South Korean Navy as shown in Table 1-3.

Table 1-3 Classification of spare parts in the South Korean Navy (Korean Navy, 2003)

Classification	Description	Example
Assembly	An end system that is assembled from components or parts	The engine or the generator of a warship
Component	The part of an end system composed of parts	The pump or the cooler of a generator
Part	The smallest unit composing an end system	The cover or the ring of a pump and a cooler

Assemblies and components denote repairable items; and parts indicate consumable items. If failed spare parts can be repaired (i.e. reparable spare parts), no further budget is required for resupply. However, reparable spare parts are rare cases. In practice, many spare parts are consumable (Rustenburg et al., 2001). This research focuses on the forecasting of consumable spare parts.

A time series is defined as "a collection of observations made sequentially through time" (Chatfield, 2004, p. 1). A stochastic process is defined as "a collection of random variables that are ordered in time and defined at a set of time points" (Chatfield, 2004, p. 33). A stochastic process is presented as a model for an observed time series (Cryer and Chan, 2008). A strictly stationary time series is defined as "one for which the probabilistic behaviour of every collection of values $\{Y_{t1}, Y_{t2}, \ldots, Y_{tn}\}$is identical to that of the time shifted set $\{Y_{t1+k}, Y_{t2+k}, \ldots, Y_{tn+k}\}$; that is, $P\{Y_{t1} \leq c_1, \ldots, Y_{tn} \leq c_n\} = P\{Y_{t1+k} \leq c_1, \ldots, Y_{tn+k} \leq c_n\}$ for all $n = 1, 2, \ldots$, all time points $t_1, t_2, \ldots, t_n$, all numbers $c_1, c_2, \ldots, c_n$, and all time shifts $k = 0, \pm1, \pm2, \ldots$" (Shumway and Stoffer, 2006, p. 23). Strict stationarity is particularly true for normal (i.e. Gaussian) process (Chatfield, 2004, Shumway and Stoffer, 2006).

Forecasting Naval spare parts is a difficult issue. This might arise from the non-normality of the spare parts demand (Willemain et al., 1994, Regattieri et al., 2005, Syntetos and Boylan, 2005). Demand that is characterised as infrequent demand occurrences of irregular demand sizes, when demand actually occurs, is an example of non-normal demand (Boylan et al., 2008). Boylan et al. (2008, p. 474) categorized non-normal demand as follows:

a) an intermittent demand item is an item with infrequent demand occurrences;
b) a slow moving item is an item whose average demand per period is low. This may be due to infrequent demand occurrences, low average demand sizes or both;
c) an erratic demand item is an item whose demand size is highly variable;
d) a lumpy demand item is an intermittent item for which demand, when it occurs, is highly variable; and
e) a clumped demand item is an intermittent item for which demand, when it occurs, is constant (or almost constant).

Non-normality of demand encompasses many demand features as shown above. Many researchers have pointed out that a large part of the time series of spare parts demand exhibits non-normal characteristics (Willemain et al., 1994, Ghobbar and Friend, 2002, Ghobbar and Friend, 2003, Willemain et al., 2004, Regattieri et al., 2005). Spare parts demand for militaries is more likely to be non-normal. Previous research about military spare parts has demonstrated that the time series of the spare parts demand are non-normal: the spare parts for helicopters in the US Army (Markland, 1970); the spare parts for the US Navy (Businger and Read, 1999); and the spare parts for the UK Air Force (Eaves and Kingsman, 2004).

Irregular large orders from a few large customers (e.g. Naval warships) can be highly sporadic. A few large customers for spare parts could induce non-normality (Silver, 1970, Eaves, 2002). The fleet of the South Korean Navy is characterised by a small number of large warships (Saunders, 2009). Therefore, the Naval spare parts demand is expected to be non-normal. Non-normal demand is difficult to forecast, because the demand occurs sporadically with some time periods showing zero demand whilst when demand does occur then the size of the demand is erratic (Willemain et al., 1994, Regattieri et al., 2005, Syntetos and Boylan, 2005). The non-normality of the Naval spare parts is described in more detail in Chapter 4.

1.2.3 Hierarchical structure

A time series for individual items is known as an item level time series. An aggregated time series for more than two items is called a group level time series. A multi-level time series structure consists of item level time series and a group level time series in which the items are members. This is known as a hierarchical structure (Hyndman et al., 2007).

A demand can be either dependent or independent. A demand is regarded as an independent demand when estimates of the demand have to be forecast; whereas a demand is regarded as a dependent demand when estimates of the demand can be calculated directly from known physical or technical relationships (DeLurgio, 1998). An item level time series might be dependent upon the group level time series (Schwarzkopf et al., 1988, DeLurgio, 1998, Fliedner, 1999, Widiarta et al., 2006). Hence, under a hierarchical structure, an aggregate pattern of demand comprised of several item level spare parts time series can be analysed using the group level time series.

In this research, a dependent demand structure for spare parts, known as the National Stock Number (NSN) which is utilised in North Atlantic Treaty Organisation (NATO) countries, is employed to take advantage of its hierarchical structure. The NSN structure is described in more detail in Chapter 4.

1.2.4 Forecasting non-normal demand

There are two forecasting strategies. A forecasting strategy which ignores the hierarchical structure of time series and simply generates a forecast at item or group level using item or group level time series is variously known as a traditional forecasting, independent forecasting, or direct forecasting (DF) (Fliedner and Lawrence, 1995, Fliedner, 1999, Miller et al., 2007). A forecasting strategy which derives a forecast at item or group level using the hierarchical structure of time series is variously known as a family-based forecasting, pyramidal forecasting, dependent forecasting, derived forecasting, or hierarchical forecasting (HF) (Fliedner and Lawrence, 1995, Fliedner, 1999, Zotteri et al., 2005, Miller et al., 2007). When an item level demand is volatile and intermittent, a higher group level demand is probably less volatile and less intermittent. This is because the volatility and intermittency of an item level demand can be offset by other item level demand in the group (Widiarta et al., 2009). This lower level of volatility and less intermittency of a group level demand could guarantee a more reliable item level demand forecast using a hierarchical forecasting strategy (Fliedner and Lawrence, 1995, Fliedner, 1999).

Top-down forecasting (TD) and bottom-up forecasting (BU) are considered to be hierarchical forecasting, because they consider the hierarchical structure of time series

(Fliedner, 1999). Top-down forecasting models a forecast at the top group level using the top group level time series of a hierarchical time series, and then creates lower level forecasts according to the item's percentage contribution within the group (Schwarzkopf et al., 1988). Bottom-up forecasting models forecasts separately for each individual item level demand, and then sums the contemporaneous item level forecasts to forecast the group level demand (Schwarzkopf et al., 1988, Fliedner, 1999).

Combinatorial forecasting (CF) models forecasts at all levels of a hierarchical time series using all levels of the time series, and then creates lower level forecasts based on a combination of the forecasts at all levels (DeLurgio, 1998, Kahn, 1998, Hyndman et al., 2007). Combinatorial forecasting can be considered to be hierarchical forecasting, because it considers the hierarchical structure of demands. Combinatorial forecasting could produce superior forecasting performance to top-down forecasting or direct forecasting (Kahn, 1998, Dekker et al., 2004, Hyndman et al., 2007).

Top-down forecasting can produce item level forecasts, bottom-up forecasting can produce group level forecasts, and combinatorial forecasting can produce both level forecasts. The major concern of this research is the hierarchical forecasting strategies which can produce item level demand (i.e. individual spare parts demand) forecasts such as top-down forecasting and combinatorial forecasting.

As mentioned above, the spare parts demand is expected to be non-normal and difficult to forecast. However, there can be some hidden features of the pattern of demand for spare parts, such as seasonality, or some other trend in the time series. An advantage of hierarchical forecasting is that it can bring out these hidden demand features so as to decrease forecasting errors (DeLurgio, 1998). A comparative analysis between alternative forecasting strategies (i.e. hierarchical forecasting and direct forecasting) is made in Chapter 5 and 6 in order to identify the superior forecasting strategy for non-normal demand.

However, many previous studies demonstrated that the relative performance of the alternative forecasting strategies (i.e. hierarchical forecasting and direct forecasting) is conditional on some specific demand features such as correlations between the time series of items within a group (Widiarta et al., 2006) and forecasting horizon (Shlifer and Wolff, 1979). Forecasting horizon is defined as "the length of time into the future for which forecasts are to be prepared" (Hyndman et al., 1998, p. 599). Owing to the conditional performance, identifying demand features which influence upon the conditional performance of the alternative forecasting strategies for the demand is important.

1.3 Aim and Objectives

The aim of this research is to establish an appropriate forecasting strategy for predicting the demand for spare parts in the South Korean Navy. The objectives of this research are to:

a) Clarify the nature of the spare parts demand in the South Korean Navy;
b) Compare the performance of the alternative forecasting strategies (i.e. top-down forecasting, combinatorial forecasting and direct forecasting) for predicting the spare parts demand at item level under the inventory control environment of the South Korean Navy;

c) Investigate the influence of demand features upon the performance of the alternative forecasting strategies; and

d) Develop a classification model for the spare parts demand in order to predict a superior forecasting method.

The main objective of this research is to compare the alternative forecasting strategies within the context of spare parts demand for the South Korean Navy. The comparisons of various direct forecasting methods are beyond the scope of this research.

1.4 Research Questions

The objectives of this research are to be achieved by answering the following research questions:

a) What is the nature of the spare parts demand in the South Korean Navy?

b) What forecasting method is appropriate for the spare parts demand in the South Korean Navy?

c) Under what conditions are top-down forecasting or combinatorial forecasting superior or inferior to direct forecasting?

d) How can the spare parts demand be classified in order to predict a superior forecasting method?

The above research questions are answered sequentially in the following chapters. Question a) is answered in Chapter 4; questions b) and c) are addressed in Chapter 5 and 6; and a solution to question d) is proposed in Chapter 6.

1.5 Research Gaps and Contributions

Hierarchical forecasting might be applicable to predicting spare parts demand in the South Korean Navy for the following reasons:

a) The data obtained from the Navy contain missing or unreliable data. Schwarzkopf et al. (1988) found that hierarchical forecasting was more accurate than direct forecasting for such data.

b) As stated, spare parts demand is expected to be non-normal. In practice there can be some hidden features in the pattern of demand for spare parts, such as seasonality, or some other trend in the time series. Hierarchical forecasting can bring out some hidden demand features, so as to decrease forecasting errors (DeLurgio, 1998).

c) A large proportion of the Naval spare parts is substitutable; that is, spare parts for specific equipment are used in a similar series of other equipment. This is because the South Korean Navy purchased a series of equipment from the same manufacturers to ensure stability of supply and continued technical support. Such substitutability makes forecasting more difficult; however, research argued that hierarchical forecasting presented more accurate forecasts for the highly substitutable demand than direct forecasting (Widiarta et al., 2008b).

d) The Naval procurement system for the spare parts requires a long forecasting horizon comprised of a long procurement lead time and long review cycle. This long forecasting horizon is a feature which could make hierarchical forecasting more accurate than direct forecasting (Shlifer and Wolff, 1979).

e) The Naval spare parts are structured by a hierarchical structure, namely the National Stock Number code. Hierarchical forecasting is an advantageous forecasting strategy for the hierarchical demand structure (Hyndman et al., 2007).

However, the literature has paid little attention to the use of hierarchical forecasting for the intermittent demand at item level, which is a feature of non-normal demand associated with spare parts demand. This is the first research gap. There is research (Fliedner and Mabert, 1992, Fliedner and Lawrence, 1995) which examined the performance of hierarchical forecasting with automotive spare parts demand. However, the authors screened out irregular time series such as time series with missing observations and demand values of zero. Viswanathan et al. (2008) examined hierarchical forecasting with intermittent demand generated by probability distributions. Nevertheless, their discussion was restricted to forecasting at group level and a production planning context with simulated data.

Second, there has been little discussion about the guidelines of hierarchical forecasting for non-normal demand. No research has investigated the influence of correlations between non-normal time series upon the relative forecasting performance of hierarchical and direct forecasting methods. Although combinatorial forecasting was argued to be a superior forecasting strategy (Kahn, 1998, Dekker et al., 2004, Hyndman et al., 2007), there has been no controlled research which has examined the influence of demand features upon the performance of combinatorial forecasting. As shown earlier, non-normal demand encompasses many demand features. In order to identify the influence of the non-normal demand features upon the performance of the forecasting strategies, a combined influence of demand features might have to be investigated. However, no research has found a combined influence of non-normal demand features upon the performance of hierarchical forecasting.

Research questions a), b), and c) are attempts to fill the first research gap. In the process of answering research question a), the non-normality expected for the spare parts demand in the South Korean Navy is identified. For research question b), the performance of hierarchical forecasting is investigated. For research question c), the forecasting conditions of the South Korean Navy and its influence upon the performance of the forecasting strategies are identified.

Research questions c) and d) are attempts to fill the second research gap. In the process of answering research question c), the influence of correlations between the time series for spare parts demand and intermittency associated with the spare parts upon the performance of combinatorial forecasting is investigated. A combined influence of multiple demand features including correlations and intermittency upon the performance of combinatorial forecasting is examined. In order to answer research question d), a multivariate classification model that predicts the relative performance of alternative forecasting methods for spare parts demand by the multivariate demand features including correlations and intermittency is proposed.

Theories on hierarchical forecasting for non-normal demand associated with spare parts demand in the South Korean Navy are not well-developed. The contributions of this research are to:

a) Identify the nature of the spare parts demand in the South Korean Navy;
b) Identify the performance of alternative forecasting strategies (hierarchical forecasting and direct forecasting) for the spare parts demand;
c) Identify the influence of demand features upon combinatorial forecasting;

d) Develop a new classification model for the spare parts demand which predicts the relative performance of the alternative forecasting methods (hierarchical and direct forecasting methods) by the multivariate demand features; and

e) Validate the research findings with diagnostics, cross-validation and a variety of accuracy measures including an accuracy measure using simulation with empirical data.

1.6 Outline of the Thesis

The remainder of this thesis presents the following chapters:

Chapter 2 reviews the literature on various topics including: forecasting methods in general, direct forecasting strategy for non-normal demand, hierarchical forecasting strategy, classification of demand for forecasting, and measures of forecasting accuracy.

Chapter 3 introduces the deployment and development of the methodology for this research. This includes: the purpose of the research, research methodologies in operations management, the choice of a case study strategy, the relationship between theory and a case study, research design, evaluating research, and the research procedure of this research.

Chapter 4 identifies the nature of the spare parts demand in the South Korean Navy containing the general information and the logistical system in the South Korean Navy, the analysis of spare parts demand data obtained from the Navy, decomposition, and the sources of non-normality.

Chapter 5 compares the performance of a range of forecasting methods with a variety of accuracy measures. This chapter comprises forecasting in the South Korean Navy, measures of forecasting accuracy including an accuracy measure using simulation, and the performance of direct and hierarchical forecasting methods.

Chapter 6 suggests a classification scheme for selecting forecasting methods. This chapter includes the competing performance of direct forecasting and hierarchical forecasting, possible demand features to guide the selection of a forecasting method, the process of classification, and classification results.

Chapter 7 presents the conclusion of this research. This chapter includes a summary of the findings with regard to the existing theories, the contributions of this research, a forecasting strategy for the South Korean Navy, limitations, and suggestions for further research.

Chapter 2. Literature Review

As mentioned in Chapter 1, the supply of Naval spare parts relies heavily on forecasting accuracy. However, identifying an appropriate forecasting method for a specific case is difficult. This is because forecasting performance is situational; that is, no single forecasting method is universally applicable (DeLurgio, 1998, Chatfield, 2004).

This chapter begins by reviewing general forecasting methods. In Section 2.2, previous research on direct forecasting method for non-normal demand is discussed. In Section 2.3, theories about hierarchical forecasting are reviewed. In Section 2.4, the comparative performance of forecasting strategies and related issues including hierarchical forecasting for non-normal demand are discussed. In Section 2.5, the classification of demand for forecasting including the influence of demand features upon forecasting performance is examined. In Section 2.6, measures of forecasting accuracy are reviewed. In Section 2.7, issues related to measuring an inventory model are examined. Finally, a summary and concluding remarks are presented in Section 2.8.

2.1 Forecasting Methods in General

There are a variety of forecasting methods which demonstrate different performance in different situations (DeLurgio, 1998, Chatfield, 2004). Hyndman et al. (1998) categorised forecasting methods as either quantitative or qualitative. They argued that a quantitative method can be used when sufficient quantitative information is available, and a qualitative method can be used when little or no quantitative information is available, but sufficient qualitative knowledge exists. This research deals solely with quantitative approaches.

Quantitative methods can be classified as either univariate or multivariate (DeLurgio, 1998). Univariate forecasting methods forecast the future based on a model fitted only to present and past observations of a given time series. In comparison, multivariate forecasting methods forecast the future based at least partly on values of one or more additional time series (Chatfield, 2004).

A forecast quantity of any time series is a single figure; however, the value of the time series is actually made up of five components (Bowersox and Closs, 1996, Silver et al., 1998): base demand component, seasonal component, trend component, cyclic component, and irregular component. The base demand component identifies the scale of a time series; the seasonal component captures seasonal variations which repeat themselves over each period; the trend component establishes the rate of growth or decline of a time series over time; the cyclic component identifies variations at a fixed period apart from the seasonal component; and the irregular component is the residuals after the other four components are identified and removed from a time series (Hyndman et al., 1998, Silver et al., 1998, Chatfield, 2004). Since the irregular component includes a random or unpredictable quantity, it makes demand forecasting difficult (Silver et al., 1998).

Bowerman et al. (2005) intimated that all forecasting situations involve some degree of uncertainty due to an irregular component in the description of a time series; that is, some errors in forecasting must be expected. No matter how well-designed or sophisticated demand forecasting is, some degree of forecasting errors cannot be avoided.

Ghobbar and Friend (2002, 2003) examined weekly, monthly and quarterly spare parts demand (e.g. battery, main undercarriage unit and brake assembly unit) for aircrafts (i.e. Fokker, BAe and ATR) in KLM-UK and identified the trend component, the seasonal component and the irregular component for a time series of the spare parts demand. They observed that a large number of items exhibit a non-normal demand pattern. The large proportion of non-normal demands of the spare parts might be caused by a large proportion of an irregular component in the times series of spare parts demands, because the trend component and the seasonal component are distinguishable from the non-normal demand patterns. Owing to the high proportion of items which exhibit non-normal demand patterns, forecasting spare parts demand is probably the biggest difficulty in the repair and overhaul industry (Ghobbar and Friend, 2002, 2003).

2.2 Direct Forecasting

As stated in Subsection 1.2.4, two forecasting strategies for non-normal demand can be recommended: a direct forecasting strategy; or a hierarchical forecasting strategy when a group of similar, related time series is identified (DeLurgio, 1998). In this section, issues related to the performance of direct forecasting methods for non-normal demand are reviewed, and then research on hierarchical forecasting methods is reviewed later in this chapter.

2.2.1 Transformation for forecasting

Transformation refers to a process of taking data which are non-normally distributed and converting them to approximate normally distributed data (Gaussian distributions)

(Miles and Shevlin, 2001). When data deviate from a normal distribution, it is sensible to consider transforming the data (Miller Jr., 1986, Chatfield, 2004).

A general class of transformation is the Box-Cox transformation (Box and Cox, 1964). A particular transformed series can be defined as in equation (2-1) (Box and Cox, 1964, p. 214). The value of the transformation parameter, λ, can be estimated by an appropriate inferential procedure (e.g. maximum likelihood). For any set of observations, $y_1, y_2, \ldots, y_n$, the likelihood function is defined as "the joint probability density of obtaining the data actually observed" (Cryer and Chan, 2008, p. 158). The maximum likelihood estimator is then defined as "the value of the parameter for which the data actually observed is most likely, that is, the value that maximise the likelihood function" (Cryer and Chan, 2008, p. 158). The estimation procedure of the best value of λ is quite complex and difficult (Nelson and Granger, 1979).

$$y_t^{(\lambda)} = \begin{cases} (y_t^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ \log y_t & \lambda = 0 \end{cases} \qquad (2\text{-}1)$$

where: $y_t^{(\lambda)}$ = the transformed value at time t

y_t = the observed value at time t

λ = transformation parameter

Some authors have claimed that the model built for the transformed data is less helpful compared to the complexity and difficulty of the Box-Cox transformation (Nelson and Granger, 1979, Chatfield, 2004). Nelson and Granger (1979) examined the forecasting performance of autoregressive integrated moving average (ARIMA) models (see Subsection 2.2.3) coupled with the Box-Cox transformation for 21 empirical time series in various contexts (e.g. steel prices, sales of grocery stores, and index of stock prices) and forms (i.e. monthly series over 245 ~ 360 observations or quarterly series over 111 ~ 115 observations). They claimed that the data were extremely non-normal and no value of λ produced normally distributed data. Using the Box-Cox transformation, the forecasting performance improvement was found to be very small (approximately 2% at most in reduction of RMSE for one-period ahead forecasting) (Nelson and Granger, 1979).

Tukey (1977) suggested that a simple transformation (i.e. replacing the raw numbers by the same simple power of each of the numbers) is useful. For example, this can be expressed as square roots (1/2 powers), reciprocals (-1 powers), or reciprocals of square roots (-1/2 powers).

Chatfield (2004) recommended avoiding transformation for the following reasons: a) skewness is difficult to eliminate with a transformation; b) it is more difficult to interpret the transformed data; and c) when the model is transformed back to be of use, the reverse transformation can cause biasing effects.

Miles and Shevlin (2001) noted that the linear transformation using a quadratic, cubic, log, or inverse function makes a case, which is close to being an outlier before transformation, an extreme case (i.e. an outlier) after transformation. The linear transformation is unlikely to handle outliers, so it does not mitigate the non-normality.

Miller Jr. (1986) conceded that the linear transformations rarely handle outliers, so he recommended robust estimators such as trimming and winsorizing which convert outliers into proximity with the rest of the data. Trimming discards small portions (e.g.

0.1 or 0.05) of each tail in the data distribution in order to remove aberrant values. Winsorizing replaces the tails by a smaller (or larger) value. As such, Businger and Read (1999) used a winsorized data set to produce forecasts for predicting the spare parts demand in the US Navy. However, it should be noted that the robust estimators are based on the assumption that underlying distribution is symmetric about its median (Miller Jr., 1986).

2.2.2 Multivariate versus univariate forecasting methods

As stated above, quantitative forecasting methods can be categorised as either univariate forecasting methods or multivariate forecasting methods. A multivariate forecasting method might be classified as a direct forecasting method, because multivariate forecasting methods do not consider the hierarchical structure of demands. While multivariate forecasting methods are expected to be at least as good as univariate forecasting methods, there are many practical cases in which a univariate forecasting method outperforms a multivariate forecasting method (Chatfield, 2004).

Although multivariate models generally generate a better fit to given data than univariate models, this supremacy does not necessarily convert into superior forecasts due to the sensitivity of multivariate models to changes in structure (Chatfield, 2004). Multivariate models can be expected to give more accurate forecasts than univariate models when a high cross-correlation between time series, such as financial time series, is observed (du Preez and Witt, 2003). Cross-correlation is defined as "a standardised measure of association between one time series and the past, present, and future values of another time series" (Hyndman et al., 1998, p. 594). However, Willemain et al. (1994) found that the sparseness of the empirical intermittent sales data of electrical equipment, jet engine tools, veterinary health, and consumer food items makes most of the cross-correlation non-significant. Therefore, multivariate models are unlikely to produce more accurate forecasts for intermittent demand than univariate models.

For the purpose of forecasting non-normal demand, univariate forecasting methods have been recognised as appropriate forecasting methods. These include exponential smoothing (ES) (Markland, 1970, DeLurgio, 1998), weighted moving average (WMA) (Ghobbar and Friend, 2003), exponentially weighted moving average (EWMA) (Regattieri et al., 2005), Croston's method (Croston, 1972), Syntetos-Boylan approximation (SBA) (Syntetos and Boylan, 2005) and Box-Jenkins models (Businger and Read, 1999).

2.2.3 Direct forecasting for non-normal demand

The standard method for forecasting non-normal demand is considered to be exponential smoothing (ES) (Croston, 1972, Sani and Kingsman, 1997, Narasimhan et al., 1998, Ghobbar and Friend, 2003). Simple exponential smoothing (SES) is the method that gives more recent observations more weight than older observations, because it is based on the thought that older demands are less related to more recent demands (Waters, 1991). The recurrence formula of simple exponential smoothing can be written as:

$$\hat{y}_t(1) = \alpha y_t + (1-\alpha)\hat{y}_{t-1}(1) \tag{2-2}$$

where:

$\hat{y}_t(1)$ = the one period ahead forecast made at time t

y_t = the observed demand for an item at time t

α = smoothing parameter ($0 < \alpha < 1$)

As α is closer to 0, the estimation is less changeable by the latest updating, whereas as α is closer to 1, the estimation is more adjusted to the latest periods. For non-normal demand, a very small smoothing parameter for simple exponential smoothing, α between 0.01 and 0.05, was considered to be appropriate because of a large amount of noise in the demand pattern (Narasimhan et al., 1998).

Exponential smoothing has some benefits: a) it hedges in that only a part of the forecast error is used by updating the $\hat{y}_{t-1}(1)$; b) it reduces arithmetic by only requiring the most recent observation to update forecasts; and c) it could save data storage by only requiring retention of the most recent $\hat{y}_{t-1}(1)$ (Silver et al., 1998, Chatfield, 2004, Sani and Kingsman, 1997). Simple exponential smoothing should generally be used for non-seasonal time series showing no systemic trend (Chatfield, 2004). Exponential smoothing was intimated to be available for a time series in which seasonality and trend are measured or removed (Waters, 1991, Gardner Jr. and Diaz-Saiz, 2002).

A seasonal component can be referred to as additive; however, when the size of a seasonal effect is directly proportional to the mean, the seasonal component is referred to as multiplicative (Chatfield, 2004). The additive seasonality model can be defined as in equation (2-3); the multiplicative seasonality model can be defined as in equation (2-4) or equation (2-5) (Chatfield, 2004, p. 20).

$$y_t = m_t + s_t + \varepsilon_t \tag{2-3}$$

$$y_t = m_t \times s_t + \varepsilon_t \tag{2-4}$$

$$y_t = m_t \times s_t \times \varepsilon_t \tag{2-5}$$

where: m_t = the deseasonalized mean level at time t

s_t = the seasonal effect at time t

ε_t = the random error at time t

For intermittent time series, it is impossible to calculate multiplicative seasonality. In this case, Gardner Jr. and Diaz-Saiz (2002) conducted an *ad hoc* procedure; that is, adding a constant before decomposition and removing it afterward.

Gardner Jr. and Diaz-Saiz (2002) compared the forecasting performance of exponential smoothing models coupled with multiplicative and additive decompositions for predicting automotive spare parts demand at BPX Holding Corporation of Huston. 691 monthly demand time series over 3 years including some intermittent time series were tested. They found that the additive seasonal models minimised forecasting errors significantly. A 20% reduction in mean absolute deviation (MAD) using the additive seasonal decomposition compared to the multiplicative seasonal decomposition was observed. They argued that additive seasonal models work well with intermittent data and are more robust to outliers.

A more complicated version of exponential smoothing such as Holt-Winters forecasting (Chatfield, 2004) is adaptable to seasonality and trend without coupling decomposition procedures. Trend and seasonality can be dealt with using Holt-Winters forecasting. The three values, L_t, T_t and I_t are updated by three smoothing parameters, α, γ and δ: the smoothing parameters are usually valued between 0 and 1. The seasonal index, I_t, can represent either additive or multiplicative: that is, $y_t - I_t$ is deseasonalised value for the additive case; whereas, y_t/I_t is for the multiplicative case. Recurrence equations for updating L_t, T_t and I_t for monthly multiplicative data and the τ periods ahead forecast made at time t can be written as (Chatfield, 2004):

$$L_t = \alpha(y_t / I_{t-12}) + (1-\alpha)(L_{t-1} + T_{t-1}) \tag{2-6}$$

$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1} \tag{2-7}$$

$$I_t = \delta(y_t / L_t) + (1-\delta)I_{t-12} \tag{2-8}$$

$$\hat{y}_t(\tau) = (L_t + \tau T_t)I_{t-12+\tau} \tag{2-9}$$

where:

L_t = local level at time t

T_t = trend at time t

I_t = seasonal index at time t

y_t = the observed demand for an item at time t

$\hat{y}_t(\tau)$ = the τ periods ahead forecast made at time t

There are various direct forecasting methods for non-normal demand. Table 2-1 compares the results of eleven major studies that have examined the forecasting performance of various direct forecasting methods for non-normal demand. In the literature the methods have been examined using empirical studies or simulation. The researchers cited obtained data from a range of sources.

Table 2-1 A review of direct forecasting methods for non-normal demand

Reference	Forecasting methods	Method of comparison	Data: Source	Data: Demand pattern	Performance
Nelson and Granger (1979)	ARIMA coupled with transformation & non-transformation	Empirical study Simulation	Various time series	Non-normal	The forecasting performance improvement is non-significant
Willemain et al. (1994)	SES & Croston	Empirical study	Demand for electrical equipment, jet engine tools, veterinary health, & food	Intermittent & highly variable	Croston presents an average of 1 ~ 14% smaller MAPE
Johnston and Boylan (1996)	Croston & EWMA	Simulation	Probability distributions		Croston presents smaller MSE when ADI is greater than 1.25 forecast review periods
Sani and Kingsman (1997)	SES, SMA & Croston	Empirical study	Spare parts demand for vehicles and machinery	Low & intermittent	SMA minimises inventory costs
Businger and Read (1999)	ES & ARIMA	Empirical study	Spare parts demand for the US Navy	Extremely volatile	ARIMA models presents 9% ~ 18% smaller forecasting errors
Gardner Jr. and Diaz-Saiz (2002)	ES coupled with additive & multiplicative decompositions	Empirical study	Automotive spare parts demand	Intermittent & seasonal	ES coupled with additive decomposition presents 20% smaller MAD
Ghobbar and Friend (2003)	13 methods including SES, Croston, WMA & EWMA	Empirical study	Spare parts demand for aircraft	Erratic, intermittent, or lumpy	WMA minimises MAPE
Eaves and Kingsman (2004)	ES, MA, Croston & SBA	Empirical study	Spare parts demand for the UK Air force	Intermittent or slow moving	SBA minimises stock-holding
Syntetos and Boylan (2005)	SES, SMA, Croston & SBA	Empirical study	Automotive spare parts demand	Slow moving, lumpy, or intermittent	SBA minimise forecasting errors
Regattieri et al. (2005)	10 methods including SES, Croston, WMA & EWMA	Empirical study	Spare parts demand for aircraft	Intermittent	WMA minimises MAD & MAD/A
Jiafu et al. (2009)	ARMA	Empirical study	Spare parts demand for power plant generating units	Slight fluctuation	Relative error: 3.14%

Key: AR(I)MA = autoregressive (integrated) moving average; (S)ES = (simple) exponential smoothing; Croston = Croston's method; EWMA = exponentially weighted moving average; (S)MA = (simple) moving average; SBA = Syntetos-Boylan approximation; ADI = average inter-demand interval; MAPE = mean absolute percentage error; MSE = mean squared error; MAD(/A) = mean absolute deviation (divided by average demand).

Some authors noted that exponential smoothing can be biased when the demand pattern is non-normal (Croston, 1972, Eaves and Kingsman, 2004, Silver et al., 1998). For instance, immediately after a transaction occurs in a period, the forecast of exponential smoothing will exceed the average demand, whereas if any transaction does not occur in a period, the forecast of exponential smoothing will be below the average.

Croston (1972) theorised Croston's method, which is argued to be an unbiased forecasting method. Croston's method can be expressed as in equation (2-10) (Eaves, 2002). Croston's method combines the estimates of the demand size (z_t) and the estimates of the demand interval (p_t) in order to estimate the mean demand per period. If demand occurs every period, Croston's method is equivalent to exponential smoothing.

If $y_t = 0$,

$p_t = p_{t-1}$; $z_t = z_{t-1}$; and $q = q + 1$

else,

$p_t = p_{t-1} + \alpha(q - p_{t-1})$; $z_t = z_{t-1} + \alpha(y_t - z_{t-1})$; and $q = 1$

$$\hat{y}_t = z_t / p_t \qquad (2\text{-}10)$$

where:

p_t = Croston's estimate of mean interval between transactions

z_t = Croston's estimate of mean demand size

q = time interval since last demand

y_t = the observed demand for an item at time t

α = smoothing parameter ($0 < \alpha < 1$)

$\hat{y}_t$ = Croston's estimate of mean demand per period

Croston's method has subsequently been corroborated by several other researchers (Willemain et al., 1994, Johnston and Boylan, 1996, Sani and Kingsman, 1997). Willemain et al. (1994) compared simple exponential smoothing and Croston's method for intermittent and highly variable demand, which is a pattern of non-normal demand, in terms of mean absolute percentage error (MAPE). 54 empirical data sets from four sources such as electrical equipment demand over 36 months, jet engine tools demand over 910 days, veterinary health items demand over 28 months, and consumer food items demand over 210 weeks were used. They argued that Croston's method presents an average of 1 ~ 14% smaller MAPE than simple exponential smoothing. Various measures of forecasting accuracy are reviewed in Section 2.6 in detail.

Johnston and Boylan (1996) compared Croston's method and exponentially weighted moving average (EWMA) with generated data [Poisson distribution for demand interval and several different probability distributions (i.e. exponential, Erlang and rectangular) for demand size]. These probability distributions are described in Appendix A. They postulated that Croston's method presents smaller mean squared error (MSE) when average inter-demand interval is greater than 1.25 forecast review periods (i.e. weeks). In exponentially weighted moving average, weightings for previous observations decrease exponentially (i.e. giving more weighting to later data). A one period ahead forecast by exponentially weighted moving average can be illustrated as in equation (2-11). If the number of moving time periods is identical to the entire data periods available, exponentially weighted moving average is equivalent to exponential smoothing.

$$\hat{y}_t(1) = \alpha y_t + (1-\alpha)\hat{y}_{t-1}(1) \qquad (2\text{-}11)$$

where:

$\hat{y}_t(1)$ = the one period ahead forecast made at time t

y_t = the observed demand for an item at time t

$\alpha = 2/(n+1)$

n = the number of the time periods

Sani and Kingsman (1997) also demonstrated the superiority of Croston's method, in that Croston's method reduces inventory costs for low and intermittent demand (i.e. 30 daily spare parts demand for vehicles and agricultural machinery over five years), which is a pattern of non-normal demand , compared to simple exponential smoothing.

There are other researchers who believe that Croston's method has more modest benefits than exponential smoothing (Eaves, 2002, Eaves and Kingsman, 2004). Croston's method appears to reduce the bias of exponential smoothing, but not perfectly. After simulation experiments with intermittent data, Syntetos and Boylan (2001) claimed that the combined ratio (z_t / p_t) of Croston's method can fail to produce accurate estimates of demand per time period. This was corroborated by several researchers (Eaves, 2002, Eaves and Kingsman, 2004, Syntetos and Boylan, 2005). These researchers argued that the bias of Croston's method increases when the smoothing parameter, a, increases: the smoothing parameter, a, greater than 0.3 introduces considerable bias. They indicated that Syntetos-Boylan approximation (SBA) provides a reasonable approximation of the actual demand per period, especially for very low volumes and large intervals between transactions. Syntetos-Boylan approximation can be written as:

$$\hat{y}_t = (1-\alpha/2)z_t/p_t \qquad (2\text{-}12)$$

where:

p_t = Croston's estimate of mean interval between transactions

z_t = Croston's estimate of mean demand size

$\hat{y}_t$ = Syntetos-Boylan approximation of mean demand per period

α = smoothing parameter ($0 < \alpha < 1$)

Eaves and Kingsman (2004) collated Syntetos-Boylan approximation with exponential smoothing, moving average and Croston's method using an inventory control model simulation (see Subsection 2.6.3) with 18,750 empirical spare parts demand data (quarterly, monthly, and weekly time series over six years), which have an intermittent or a slow moving demand pattern, in the UK Air Force. They showed that an inventory system using SBA reduces stock-holding significantly compared to other methods. Syntetos and Boylan (2005) also verified that SBA is the most accurate forecasting method with 3,000 demand data (monthly time series over two years) from the automotive industry, which have a slow moving, a lumpy, or an intermittent demand pattern, when compared to simple exponential smoothing (SES), simple moving average (SMA), and Croston's method.

Simple moving average (SMA) is a forecasting model based on the last n demand figures and ignores older values. Simple moving average can be expressed as:

$$\hat{y}_t(1) = \frac{1}{n}\sum_{i=1}^{n} y_{t-i+1} \tag{2-13}$$

where:

$\hat{y}_t(1)$ = the one period ahead forecast made at time t

y_t = the observed demand for an item at time t

n = the number of the time periods

Simple moving average has some advantages and disadvantages (Bowersox and Closs, 1996, Choi et al., 2005, Silver et al., 1998). In terms of advantages, a) it can be calculated easily; and b) it is able to reflect the latest trend. The disadvantages are that, a) a large amount of data must be maintained; and b) it is unresponsive or sluggish to change, because it does not consider seasonal change, and equal weight is allocated to the n most recent pieces of data. Sani and Kingsman (1997) argued that simple moving average minimises inventory costs for the low and intermittent demand (i.e. 30 daily spare parts demand for vehicles and agricultural machinery over five years), compared with Croston's method and simple exponential smoothing.

More immediate past observations might be more relevant in forecasting than older observations. While simple moving average places equal weight to each observation, weighted moving average (WMA) gives more weight to the more recent data. Weighted moving average can be expressed as:

$$\hat{y}_t(1) = \sum_{i=1}^{n}(n-i+1)y_{t-i+1} \Big/ \sum_{i=1}^{n} n \tag{2-14}$$

where:

$\hat{y}_t(1)$ = the one period ahead forecast made at time t

y_t = the observed demand for an item at time t

n = the number of the time periods

Weighted moving average was argued to be a superior forecasting method for predicting spare parts demand for airline fleets when compared to other methods such as exponential smoothing and Croston's method (Ghobbar and Friend, 2003, Regattieri et al., 2005). Ghobbar and Friend (2003) found that most of the weekly, monthly and quarterly spare parts (e.g. battery, main undercarriage unit and brake assembly unit) demand for aircrafts (i.e. Fokker, BAe and ATR) in KLM-UK are erratic, intermittent, or lumpy. Then, they compared thirteen different methods (including simple exponential smoothing, Croston's method, weighted moving average, and exponentially weighted moving average) in terms of MAPE and argued that weighted moving average is superior to other methods.

Regattieri et al. (2005) corroborated the superiority of weighted moving average. They compared 10 forecasting methods including weighted moving average, simple exponential smoothing, Croston's method and exponentially weighted moving average for predicting monthly spare parts demand for Airbus A320 aircraft of Alitalia, which has an intermittent demand pattern. Then, they argued that weighted moving average produced the most accurate results in terms of mean absolute deviation (MAD) and MAD divided by average demand (MAD/A).

An advantage of weighted moving average and exponentially weighted moving average is that the resulting smoothed trend is smoother than simple moving average. Different from simple moving average, the weights can be slowly down-weighted (Hyndman et al., 1998). Various schemes for selecting appropriate weights are available for weighted moving average and exponentially weighted moving average (DeLurgio, 1998, Hyndman et al., 1998). Weighted moving average and exponentially weighted moving average have the same disadvantages as those of the simple moving average: a) they cannot model seasonality or trend; and b) they have to store and process a large amount of data; however, it seems to be no longer a matter of concern with the improvement of computer technology (DeLurgio, 1998).

An autoregressive integrated moving average (ARIMA) model, sometimes called the Box-Jenkins model, is a tool for understanding and predicting future values. The model includes an autoregressive (AR) part in which the current value is dependent upon past values to the p^{th} depth as shown in equation (2-15) and a moving average (MA) part in which the current value is dependent upon forecast errors of previous periods as shown in equation (2-16). A model with both the autoregressive and moving average parts is usually referred to as an autoregressive moving average model [ARMA (p, q)] where p is the order of the autoregressive part and q is the order of the moving average part as shown in equation (2-17). If a time series is non-stationary, the general approach is to difference the time series using the difference operator as shown in equation (2-18). This model is normally referred to as an ARIMA (p, d, q) model where p, d, and q are integers greater than or equal to zero and refer to the order of the autoregressive, integrated, and moving average parts of the model respectively.

$$\text{AR}\,(p)\text{:}\quad y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + z_t \tag{2-15}$$

$$\text{MA}\,(q)\text{:}\quad y_t = z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q} \tag{2-16}$$

$$\text{ARMA}\,(p, q)\text{:}\quad y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q} \tag{2-17}$$

$$\text{ARIMA}\,(p, d, q)\text{:}\quad w_t = \phi_1 w_{t-1} + \cdots + \phi_p w_{t-p} + z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q} \tag{2-18}$$

where:

y_t = a value at time t

z_t = a random variable which are mutually independent and identically distributed at time t

ϕ's, θ's, p, q, and d = constants

B^d= the backward shift operator such that $B^d y_t = y_{t-d}$

∇^d = the difference operator such that $\nabla^d y_t = y_t - y_{t-d}$

$w_t = \nabla^d y_t = (1 - B)^d y_t$

When p, d, q, ϕ's and θ's are chosen and estimated, a τ periods ahead forecast made at time t can be produced by substituting $t+\tau$ for t in equation (2-18). Future error terms, z_t's, are predicted by their mean (i.e. zero).

Businger and Read (1999) argued that ARIMA models are useful in forecasting demand for spare parts in the US Navy. Approximately 12,000 quarterly spare parts demand time series having at least 20 positive demands in 40 quarters were used for their investigation. The time series were found to be extremely volatile. They compared one period ahead forecasting accuracy between exponential smoothing and ARIMA models, and found 9%, 15% and 18% smaller forecasting errors in ARIMA(1, 1, 1), ARIMA(2, 2, 2) and ARIMA(3, 2, 3) respectively.

Recently, Jiafu et al. (2009) have examined ARMA models for forecasting one year ahead demand for four spare parts (i.e. whorl, valve, atmoseal and shaft) for power plant generating units. Quarterly time series demand data for the spare parts between 2000 and 2007 were used for their examination. The time series were characterised as slight fluctuations. They argued that ARMA models for the four spare parts demands based on data between 2000 and 2006 presented reasonable forecasts; that is, the models generated 3.14% under-forecasts in 2007 (i.e. observed demand = 12,665; forecast demand = 12,267).

2.3 Hierarchical Forecasting

Direct forecasting strategy, as discussed above, ignores the hierarchical structure of a time series and simply generates forecasts at an item level. Traditionally, various direct forecasting methods were considered as appropriate forecasting methods for non-normal demand. However, as stated in Section 1.5, hierarchical forecasting might be applicable to predicting military spare parts demand which is likely to be non-normal. In this section, theories related to hierarchical forecasting strategy, which consider hierarchical demand structure, are reviewed.

2.3.1 Concept of hierarchical forecasting

Hierarchical forecasting is based on the premise that: a) aggregated time series will be less volatile and will provide better forecast performance; b) hierarchical forecasting can generate different hierarchical levels of forecast for the different hierarchical departments of an organisation; and c) hierarchical forecasting could reduce forecasting burdens by forecasting only the aggregated level demand because the individual item level demand can be produced by a simple proration method (Gross and Sohl, 1990, Fliedner and Lawrence, 1995, Fliedner, 1999, Fliedner, 2001, Widiarta et al., 2009). The first analysis using hierarchical forecasting may be traced back to the work of Theil (1954). He intimated two forecasting strategies; that is, a direct forecasting strategy and a derived forecasting (DRF) strategy. Shlifer and Wolff (1979) corroborated the two forecasting strategies: direct forecasting and derived forecasting (i.e. hierarchical forecasting). In hierarchical forecasting strategies, there are two sub-strategies, which are a top-down (TD) strategy and a bottom-up (BU) strategy (Shlifer and Wolff, 1979, DeLurgio, 1998, Fliedner and Lawrence, 1995, Fliedner, 1999, Fliedner, 2001). The two sub-strategies are described as follows.

2.3.2 Bottom-up and top-down

In a bottom-up strategy, forecasts for a group level are produced by a two-step process. In the current period, *t*, the group (or aggregated) level time series is expressed as:

$$Y_t = \sum_{i=1}^{N} y_{i,t} \tag{2-19}$$

where:

Y_t = the aggregate demand for a group of N items at time t

$y_{i,t}$ = the demand of item i at time t

$i = 1, 2, \ldots. N$

$t = 1, 2, \ldots, n$

First, the direct forecast of item i, τ periods ahead made at time t, $\tilde{f}_{i,t+\tau}$, is generated; then, the derived forecast at group level, τ periods ahead made at time t, $F_{t+\tau}$, is determined as a contemporaneous sum of the item level forecasts.

$$F_{t+\tau} = \sum_{i=1}^{N} \tilde{f}_{i,t+\tau} \tag{2-20}$$

Top-down forecasting is originated from the notion that the sum of errors in several individual item level forecasts are normally greater than the errors of the cumulative group level forecast in which the items are members (DeLurgio, 1998, Schwarzkopf et al., 1988). Following the top-down forecasting strategy, a forecast for a group level or an item level is generated by a two-step process. First, a direct forecast of demand at the group level is generated; second, the forecast is prorated to accomplish item level forecasts with a proration method as described below.

2.3.3 Proration methods

There are many possible proration methods (Gross and Sohl, 1990, Fliedner and Lawrence, 1995, DeLurgio, 1998, Narasimhan et al., 1998, Fliedner, 1999, Fliedner, 2001, Widiarta et al., 2008b). Gross and Sohl (1990) suggested a mean proportion of a given product's demand to the total product line's demands as in equation (2-21); similarly, another proportion is the proportion of a given product's mean demand to the overall product line's mean demand as in equation (2-22). In both methods, the item forecasts are produced by multiplying the group level direct forecast by the ratio of the respective item level demand.

$$f_{i,t+\tau} = \widetilde{F}_{t+\tau} \times \sum_{t=1}^{n} \frac{y_{i,t}}{Y_t} / n \qquad (2\text{-}21)$$

$$f_{i,t+\tau} = \widetilde{F}_{t+\tau} \times \frac{\sum_{t=1}^{n} y_{i,t}}{n} / \frac{\sum_{t=1}^{n} Y_t}{n} \qquad (2\text{-}22)$$

where:

$f_{i,t+\tau}$ = the hierarchical forecast of item i, τ periods ahead made at time t

$\widetilde{F}_{t+\tau}$ = the group level direct forecast, τ periods ahead made at time t

Y_t = the aggregate demand for a group of N at time t

$y_{i,t}$ = the demand of item i at time t

$i = 1, 2, \ldots. N$

$t = 1, 2, \ldots, n$

There is another mean proportion process as in equation (2-23) (Fliedner and Lawrence, 1995, Fliedner, 1999, Fliedner, 2001). An item level forecast can be calculated by multiplying the aggregated group level direct forecast by the ratio of the direct item level forecast divided by the sum of the direct forecasts at item level constituting their group.

$$f_{i,t+\tau} = \widetilde{F}_{t+\tau}(\widetilde{f}_{i,t+\tau} / \sum_{i=1}^{N} \widetilde{f}_{i,t+\tau}) \tag{2-23}$$

where:

$f_{i,t+\tau}$ = the hierarchical forecast of item i, τ periods ahead made at time t

$\widetilde{f}_{i,t+\tau}$ = the direct forecast of item i, τ periods ahead made at time t

$\widetilde{F}_{t+\tau}$ = the group level direct forecast, τ periods ahead made at time t

Gross and Sohl (1990) proposed several correlation processes using differential weights. In equation (2-24), lagged proportions (lag two, three and four, respectively) are combined using weights based on correlations. k denotes the lag of w_k; j denotes a lag to be compared; l denotes the total number of lags compared. The weights are generated by each lag's correlation divided by the sum of all correlations. Therefore, the highest correlation generates the maximum weight, the next highest the next largest weight, and so forth.

$$f_{i,t+1} = \widetilde{F}_{t+1} \times \sum_{k=1}^{l} w_k PR_{i,t-k}, \; l = 2, 3, 4 \tag{2-24}$$

where:

$$PR_{i,t} = {y_{i,t}}/{Y_t}$$

$$w_k = \frac{corr(PR_{i,t}, PR_{i,t-k})}{\sum_{j=1}^{l} corr(PR_{i,t}, PR_{i,t-j})}$$

Furthermore, Gross and Sohl (1990) extended the correlation process to the combination of higher correlations with the current period's lag. For example, equation (2-25) combines two lags having the highest correlations with the current period's lag. q denotes a lag which has the highest correlation with the current period's lag; s denotes a lag which has the second highest correlation with the current period's lag; k and l denote q or s; r denotes a lag to be compared; n denotes the total number of lags compared.

$$f_{i,t+1} = \widetilde{F}_{t+1} \times \sum_{\substack{k=1 \\ l=q,s}}^{2} w_k PR_{i,t-l} \tag{2-25}$$

$$PR_{i,t} = {y_{i,t}}/{Y_t}$$

where:

$$w_1 = corr(PR_{i,t}, PR_{i,t-q}) / \sum_{j=q,s} corr(PR_{i,t}, PR_{i,t-j})$$

$$w_2 = corr(PR_{i,t}, PR_{i,t-s}) / \sum_{j=q,s} corr(PR_{i,t}, PR_{i,t-j})$$

$$\max_{r=1 \ to \ n} \left[corr(PR_{i,t}, PR_{i,t-r})\right] = corr(PR_{i,t}, PR_{i,t-q})$$

$$\max_{\substack{r=1 \ to \ n \\ r \neq q}} \left[corr(PR_{i,t}, PR_{i,t-r})\right] = corr(PR_{i,t}, PR_{i,t-s})$$

In their empirical study with the eighteen monthly sales time series of three industrial galvanised steel products from a particular company (USG Industries, Inc.), Gross and Sohl (1990) argued that proration methods adopting equations (2-21), (2-22), and (2-25) generated relatively accurate forecasts among the various proration methods tested. Particularly, they intimated that the simple average methods [i.e. equation (2-21) and (2-22)] are uncomplicated to calculate, implement and update, and intuitively reasonable.

On the other hand, there is a proration method which uses simple exponential smoothing as shown in equation (2-26) (Narasimhan et al., 1998). $\widetilde{F}_{(t-1)+1}$ (or $\widetilde{f}_{i,(t-1)+1}$) denotes a group level direct forecast (or a direct forecast of item i), 1 period ahead made at time t-1. Widiarta et al. (2008b) argued that an accurate product's proportion in top-down forecasting could be achieved by this method because this method considers the most recent distribution of demands by giving later product's proportions more weights than older proportions of forecasts.

$$f_{i,t+1} = \left[\hat{\alpha} \frac{y_{i,t-1}}{\sum_{i=1}^{N} y_{i,t-1}} + (1-\hat{\alpha}) \frac{\tilde{f}_{i,(t-1)+1}}{\tilde{F}_{(t-1)+1}} \right] \tilde{F}_{t+1} \tag{2-26}$$

where:

$f_{i,t+1}$ = the hierarchical forecast of item i, one period ahead made at time t

$\tilde{f}_{i,t+1}$ = the direct forecast of item i, one period ahead made at time t

$\tilde{F}_{t+1}$ = the group level direct forecast, one period ahead made at time t

$y_{i,t}$ = the demand of item i at time t

$\hat{\alpha} = \min(D_t / D,\ 0.20)$

D_t = the group's total demand during period t

D = the group's total demand during last year

When one forecasting model does not greatly dominate the other forecasting model, combining forecasts of the two models was suggested to improve forecast accuracy compared to the forecast accuracy of the two models individually (DeLurgio, 1998). This combining method refers to combinatorial forecasting as mentioned in Subsection 1.2.4. DeLurgio (1998) illustrated three proration methods for combinatorial forecasting: simple averages [equation (2-27)] and two weighted averages [equation (2-28)]: weights inversely proportionate to the sum of squared errors [equation (2-30)]; and weights determined by regression analysis [equation (2-31)]. He postulated that simple averaging usually performs better than the other methods, and weights inversely proportionate to the sum of squared errors (SSE) is often better than weights determined by regression analysis. SSE can be expressed as in equation (2-29).

$$f_{i,t+\tau} = \frac{1}{2}(\widetilde{F}_{t+\tau} + \sum_{i=1}^{N} \widetilde{f}_{i,t+\tau}) \times \frac{\widetilde{f}_{i,t+\tau}}{\sum_{i=1}^{N} \widetilde{f}_{i,t+\tau}} \tag{2-27}$$

$$f_{i,t+\tau} = (w_1 \widetilde{F}_{t+\tau} + w_2 \sum_{i=1}^{N} \widetilde{f}_{i,t+\tau}) \times \frac{\widetilde{f}_{i,t+\tau}}{\sum_{i=1}^{N} \widetilde{f}_{i,t+\tau}} \tag{2-28}$$

where:

$f_{i,t+\tau}$ = the hierarchical forecast of item i, τ periods ahead made at time t

$\widetilde{f}_{i,t+\tau}$ = the direct forecast of item i, τ periods ahead made at time t

$\widetilde{F}_{t+\tau}$ = the group level direct forecast, τ periods ahead made at time t

w_i = weights of individual forecasts; $\Sigma w_i = 1.0$

Since direct forecasting models with lower squared errors can be expected to be more accurate than other direct forecasting models (although this is not so in all cases), the weights for individual forecasts should be inversely related to the squared errors. Thus, the SSE set rational weights in a way so that a higher weight is given to the more accurate forecasting model.

$$SSE = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \tag{2-29}$$

where:

y_t = the observed demand at time t

$\hat{y}_t$ = the estimated demand of y_t

$t = 1, 2, 3, \ldots, n$

n = the total number of time periods

$$w_1 = \frac{1/SSE1}{1/SSE1 + 1/SSE2}$$
$$w_2 = \frac{1/SSE2}{1/SSE1 + 1/SSE2} \tag{2-30}$$

where:

$SSE1$ = the sum of squared errors for the group level direct forecast ($\widetilde{F}_{t+\tau}$)

$SSE2$ = the sum of squared errors for all the item level direct forecasts within the group ($\sum_{i=1}^{N} \widetilde{f}_{i,t+\tau}$)

The concept of regression analysis is also used as a combining method as shown in equation (2-31).

$$f_{i,t+\tau} = (b_1 \widetilde{F}_{t+\tau} + b_2 \sum_{i=1}^{N} \widetilde{f}_{i,t+\tau}) \times \frac{\widetilde{f}_{i,t+\tau}}{\sum_{i=1}^{N} \widetilde{f}_{i,t+\tau}} \tag{2-31}$$

where: b_1 and b_2 = the regression coefficients of individual forecasts

Least square method estimates the regression coefficients (b_1 and b_2) using equation $Y_{t+\tau} = b_1 \widetilde{F}_{t+\tau} + b_2 \sum_{i=1}^{N} \widetilde{f}_{i,t+\tau}$ where the actual value, $Y_{t+\tau}$ (i.e. the aggregate demand for a group of N at time $t+\tau$), is the dependent variable and $\widetilde{F}_{t+\tau}$ and $\sum_{i=1}^{N} \widetilde{f}_{i,t+\tau}$ are independent variables.

In addition, Schwarzkopf et al. (1988) illustrated the proration updating cycle. When a demand for an item within a group is volatile, the ratio might require regular updating, probably at each forecast period. However, in a stable environment the ratio might be updated annually.

2.4 Performance of Forecasting

Several studies have discussed the comparative analysis of the performance of the forecasting strategies. Table 2-2 compares the results of 14 major studies that have compared the performance of different forecasting strategies that have used top-down forecasting, direct forecasting and combinatorial forecasting. The methods were compared using analytical models, simulation, empirical studies or some combination of these approaches. Analytical studies compare forecasting performance in terms of the variance of forecasting errors. Simulation and empirical studies compare forecasting performance in terms of the magnitude of forecasting errors. The researchers obtained data from a range of contexts and sources which present various patterns as shown in Table 2-2. The number of items or levels in the group are also various.

The research literature on the comparisons of forecasting strategies can be divided into three streams based on forecasting levels: a) group level; b) item level; and b) both levels. As stated in Section 1.3, the major concern of this research is the item level forecasting strategies. However, the group level forecasting strategy (i.e. bottom-up forecasting) is worth reviewing because it might have implications for the item level forecasting strategies.

Gross and Sohl (1990) contended that the applicability of top-down forecasting should be considered as a function of three conceptual factors: a) the group level forecasting accuracy; b) the proration accuracy; and c) the item level forecasting accuracy. As the first factor, the magnitude and the direction of group level forecasting errors are considered. When the direct forecasting at group level is accurate, the top-down forecasting can be considered. This is because an accurate item level forecast can be derived by an appropriate proration method from an accurate direct forecasting at group level. When the direct forecasting at group level is consistently under- or over-estimated, as the second factor, the proration is claimed to be biased toward the opposite direction of group level forecasting to compensate the group level forecasting errors. However, if the direct forecasting at group level is inaccurate with no eminent direction for the bias, the top-down forecasting is unlikely to produce an accurate forecast at time level. Also, when the performance of a direct forecasting method at item level is unsatisfactory, compared with the performance of a direct forecasting method at group level, the top-down forecasting method is considered.

Therefore, higher forecasting accuracy and a consistent direction of forecasting at group level might have implications for the capability of the top-down forecasting to produce higher forecasting accuracy at item level with an appropriate proration method. As such, all three streams (i.e. forecasting strategies for group level, item level, and both levels) of literature are reviewed as follows.

Table 2-2 A review of forecasting strategies and forecasting performance

Reference	Forecasting level	Forecasting strategy	Method of comparison	Context	Data		Group		Forecasting method	Forecasting performance (superior forecasting strategy or forecasting method)
					Source	Demand pattern	No. of items	No. of levels		
Shlifer and Wolff (1979)	Group Item	BU, DF & TD	Analytic study	Sales in market segment			Several	2		BU outperforms DF at group level DF outperforms TD at item level; TD for long forecasting horizon
Schwarzkopf et al. (1988)	Item	TD & DF	Analytic study	Product line			2	2	ES	No superiority
Gross and Sohl (1990)	Item	TD & DF	Empirical study	Steel product line	Steel sales data	Low r	3~7	2	ES Regression Holt-winters	DF outperforms TD in 98.4% of cases
Dangerfield and Morris (1992)	Item	TD & DF	Empirical study	Product line	M-competition data	$-0.96<r<1.0$ Various seasonal & trend patterns	2	2	ES	DF outperforms TD in 73% of cases
Fliedner and Lawrence (1995)	Item	TD & DF	Empirical study	Spare parts distribution	Demand for automotive spare parts	Erratic	95 ~ 477	3	MA SES Holt-Winters	SES DF presents 1.71% smaller MPE
Kahn (1998)	Group Item	TD, BU & CF	Empirical study	Sales	Sales data	Various seasonal patterns	14	3	ES	CF
Fliedner (1999)	Group	DF & BU	Simulation	Product line	MA (1)		2	2	SES & MA	DF using SES presents 15.6% & 29.5% smaller MAPE than DF using MA & BU using SES

Key: CF = combinatorial forecasting; DF = direct forecasting; TD = top-down forecasting; BU = bottom-up forecasting; (S)ES = (simple) exponential smoothing; MA = moving average; r = correlations among items in a group; MPE = mean percentage error; MAPE = mean absolute percentage error.

(Continued)

Table 2-2 Continued

Reference	Forecasting level	Forecasting strategy	Method of comparison	Context	Data		Group		Forecasting method	Forecasting performance (superior forecasting strategy or forecasting method)
					Source	Demand pattern	No. of items	No. of levels		
Dekker et al. (2004)	Item	DF, TD & CF	Empirical study	Wholesale (drinks & tubes)	Sales data	2 different seasonal patterns	13~29	2	Holt-Winters SES	SES CF presents 3.5-5.0% & 7.3-12.0% smaller MAD than DF & TD
Widiarta et al. (2006)	Item	TD & DF	Analytic study & Simulation	Product line	AR (1)		2	2	ES	Performance depends on lag-1 autocorrelation of the time series
Hyndman et al. (2007)	Four levels	TD, BU & CF	Simulation & Empirical study	Tourist arrivals	ARIMA Australian tourism data	Various seasonal & trend patterns	56	4	ES	CF presents 0.07% & 8.59% smaller MAPE than DF & TD at bottom level
Widiarta et al. (2008a)	Item	TD & DF	Analytic study & Simulation	Product line	MA (1)		2	2	SES	No superiority
Widiarta et al. (2008b)	Group Item	TD, BU & DF	Simulation	Product line	AR (1), MA (1), & ARMA (1, 1)		2	2	SES	DF outperforms BU for all AR(1) & ARMA(1,1) at group level Performance at item level depends on the degree of product substitutability
Viswanathan et al. (2008)	Group	DF & BU	Simulation	Product line	Probability distribution	Various intervals & demand volumes	2, 4, 6, 8, & 10	2	SES & Croston's method	BU using Croston's method when the variability of interval is low DF using SES when the variability of interval and demand size is high
Widiarta et al. (2009)	Group	DF & BU	Analytic study & Simulation	Product line	MA (1)		2	2	SES	No superiority

Key: CF = combinatorial forecasting; DF = direct forecasting; TD = top-down forecasting; BU = bottom-up forecasting; (S)ES = (simple) exponential smoothing; MA = moving average; MAD = mean absolute deviation; MAPE = mean absolute percentage error.

2.4.1 Bottom-up forecasting

Preliminary work on the performance of hierarchical forecasting was undertaken by Shlifer and Wolff (1979). In their analytic study, Shlifer and Wolff (1979) compared direct and bottom-up forecasting in terms of the variance of forecasting errors. After a series of arithmetic processes, they concluded that, at group level, bottom-up forecasting is superior to direct forecasting.

On the contrary, there are researchers (Fliedner, 1999, Widiarta et al., 2008b) who postulated that direct forecasting is superior to bottom-up forecasting. Widiarta et al. (2008b) argued that, at group level, direct forecasting outperforms bottom-up forecasting after a simulation experiment. Simple exponential smoothing was used as the forecasting method under both bottom-up and direct forecasting. Direct forecasting was argued to be superior to bottom-up forecasting in all the cases with data generated from AR (1) and ARMA (1, 1) processes and 87.5% of cases with data generated from MA (1) process (Widiarta et al., 2008b).

After his simulation study with data made by a time series generator using MA(1) process, Fliedner (1999) argued that, at group level, direct forecasting using simple exponential smoothing outperforms direct forecasting using moving average and bottom-up forecasting using simple exponential smoothing.

Some authors (Fliedner, 1999, Viswanathan et al., 2008) contended that the performance of bottom-up and direct forecasting at group level is dependent upon demand features such as correlations and the variability of demand interval and demand size. Fliedner (1999) argued that, when the correlations (regardless of whether they are negative or positive) between two time series are high, the accuracy of both direct and bottom-up forecasting at group level improves. Viswanathan et al. (2008) examined group level forecasting performance with simulation using data generated by a variety of probability distributions. The variation in item level demand size was modelled as a normal distribution, lognormal distribution or gamma distribution. The intermittent demand interval was modelled as a uniform or gamma distribution (see Table A-1 for these probability distributions). Simple exponential smoothing and Croston's method were used as the forecasting methods under both bottom-up and direct forecasting. They indicated that forecasting performance depends on the variability of demand interval and demand size; that is, when the variability of interval is low, bottom-up forecasting is superior. Conversely, when the variability of interval and demand size is high, direct forecasting is superior.

On the other hand, Widiarta et al. (2009) argued that the difference in the performance of bottom-up and direct forecasting at group level is non-significant regardless of correlations. This argument was derived from their combination of analytic study and simulation with data generated using MA (1) process. Simple exponential smoothing was used as the direct forecasting method for both bottom-up and direct forecasting.

2.4.2 Top-down forecasting

Inconsistent with the intuitive notion, top-down forecasting at item level has been observed to provide rather inferior performance (Shlifer and Wolff, 1979, Gross and Sohl, 1990, Dangerfield and Morris, 1992, Fliedner and Lawrence, 1995) or identical performance (Widiarta et al., 2008a) to direct forecasting. Shlifer and Wolff (1979) compared direct and top-down forecasting in terms of the variance of forecasting errors.

In the analytic study, they concluded that, at item level, direct forecasting is superior to top-down forecasting.

Dangerfield and Morris (1992) investigated the effects of correlations between items in a group and an item's proportion of a group on the performance of forecasting strategies with 178 monthly time series from M-competition data. Makridakis, et al. (1982) proposed 1,001 time series, known as M-competition, that are used to compare forecasting methods. 89 group level time series (each containing two items) were constructed using the 178 item level time series. The pairs of item level time series were observed to be varied with regard to correlations ($-0.96 < r < 1.0$), to seasonal and trend patterns, and to their relative proportion of the group level time series. Forecasts both at group and item levels were generated by exponential smoothing. Equation (2-21) was employed as the proration method. Dangerfield and Morris (1992) argued that direct forecasting is superior to top-down forecasting in almost three out of four data sets regardless of the correlations and the proportions.

Gross and Sohl (1990) collated top-down and direct forecasting methods with the monthly time series consisting of 53 observations for three industrial galvanised steel product sales (grip strut, channel and accessories) from USG Industries Inc. The grip strut group consisted of seven items; the channel group consisted of three items; and the accessories group consisted of five items. Several forecasting methods such as simple exponential smoothing, Holt-Winters forecasting, and linear time series regression were used for both top-down and direct forecasting methods. Several proration methods such as equation (2-21), equation (2-22), equation (2-24) and equation (2-25) were used for top-down forecasting. Very low correlations were observed among all time series, both within and between the product lines. In their study, direct forecasting was argued to outperform top-down forecasting in 98.4% of cases.

Fliedner and Lawrence (1995) compared top-down and direct forecasting methods with 954 monthly time series consisting of 42 observations for automotive spare parts demand from Cummins Engine Inc. Cummins Engine Inc. uses a hierarchical structure for classifying products; namely standard product classification. They screened out irregular time series such as time series with missing observations and demand values of zero. However, the volatile demand feature was still observed from the data sets which remained after the screening process. Among the three forecasting methods (i.e. simple exponential smoothing, Holt-Winters forecasting and moving average) simple exponential smoothing was used for both top-down and direct forecasting methods because simple exponential smoothing was argued to make the performance of top-down forecasting better. Equation (2-23) was used as the proration method. Fliedner and Lawrence (1995) argued that direct forecasting is superior to top-down forecasting. The mean percentage errors (MPE) of direct and top down forecasting methods were -74.89 and -76.60 respectively with p-value, 0.021.

Widiarta et al. (2008a) contended that no superiority is observed between top-down and direct forecasting methods. The data were generated by MA (1) process. The number of items in a group was restricted to two items. Simple exponential smoothing was employed as the forecasting method for top-down and direct forecasting methods. Equation (2-21) was used as the proration method for top-down forecasting. They argued that the performances of top-down and direct forecasting are identical in an analytic evaluation and the forecasting error difference of each forecasting strategy is non-significant in a simulation.

In practice most studies demonstrated that the performance of top-down forecasting method is conditional on some specific demand features: correlations between time series (Schwarzkopf et al., 1988, Widiarta et al., 2006); the lag-1 autocorrelation of the time series (Widiarta et al., 2006); the variability of demand (Schwarzkopf et al., 1988); forecasting horizon (Shlifer and Wolff, 1979); and the degree of substitutability and the variability of an item's proportion (Widiarta et al., 2008b). These will be discussed later in this chapter.

2.4.3 Combinatorial forecasting

Some authors (Kahn, 1998, Dekker et al., 2004, Hyndman et al., 2007) contended that combinatorial forecasting is superior to top-down forecasting and direct forecasting. Kahn (1998) claimed that combinatorial forecasting outperforms top-down forecasting and direct forecasting after an empirical study with sales data which have various seasonal patterns. Dekker et al. (2004) compared five forecasting methods: simple exponential smoothing, Holt-Winters, Holt-Winters combined with Naive 1, aggregation method, and aggregation method combined with Naive 1. Naive 1 refers to a forecast identical with the most recent demand. The aggregation method lays the foundation on the multiplicative Holt-Winters forecasting: firstly, seasonal indices at the group level are produced; then, item level demands are forecast using seasonal indices of the group level. Through experiments using five years of weekly empirical sales data (i.e. sales for 14 beers, 42 soft drinks and 11 plastic tubes for construction), they argued that the aggregation method combined with Naive 1 generated the most accurate forecasts. In their view, the reason for its superiority was that, if in a particular period the demand is too high or too low, the aggregation method combined with Naive will correct the irregularity and the forecast will remain closer to the most recent demand.

Hyndman et al. (2007) argued for the superiority of combinatorial forecasting by simulation with data from ARIMA process and an empirical study of Australian tourism. Quarterly time series on the number of visitor nights covering the period of 1998 ~ 2006 from the National Visitor Survey were used for the empirical study. These data were composed of four levels of a hierarchical structure: level 0 (top) indicates aggregate domestic tourism demand for the whole of 'Australia'; level 1 is divided by 'purpose of travel'; level 2 is divided by 'states and territories'; and level 3 (bottom) is divided by 'capital city versus other'. Exponential smoothing was used as the forecasting method for all forecasting strategies (direct forecasting, top-down forecasting, bottom-up forecasting, and combinatorial forecasting). Equation (2-20) was used for bottom-up forecasting; equation (2-21) was used for top-down forecasting. In order to produce combinatorial forecasting, all four levels were forecast individually using exponential smoothing and then these forecasts were optimally combined using a regression model. They argued for the superiority of combinatorial forecasting; that is, the combinatorial forecasting was the most superior at intermediate levels (i.e. level 1 and 2) with the simulation experiment, and the most superior at the intermediate level (i.e. level 1) and the bottom level (i.e. level 3) for the Australian tourism data. For example, with the tourism data the combinatorial forecasting method was observed to present 0.07% and 8.59% smaller mean absolute percentage error (MAPE) than those of direct forecasting and top-down forecasting respectively at the bottom level.

2.4.4 Disagreement in the performance

Lack of agreement in the performance of hierarchical forecasting methods was shown in the previously mentioned research. The disagreement could be attributed to different features of data in each investigation, for example, variations in the number of items in a group or the sources of data used as shown in Table 2-2. Fliedner (1999) noted the number of items in the groups as a source of incongruity in the literature. In an analytical study, the number of items in a group was normally constrained to two items in order to control the statistical correlation (Shlifer and Wolff, 1979, Wei and Abraham, 1981, Schwarzkopf et al., 1988, Widiarta et al., 2006, Widiarta et al., 2008a). However, in empirical studies, greater numbers of items were usually used without measuring the association of homogeneity within family groups (Fliedner and Lawrence, 1995, Kahn, 1998, Hyndman et al., 2007). Fliedner (1999) pointed out that the statistical features of data used could be another reason for incongruity. For example, a specific data generator in analytic studies (Viswanathan et al., 2008, Fliedner, 1999, Widiarta et al., 2006, Widiarta et al., 2008a) could lead to different forecast results.

2.4.5 Hierarchical forecasting for non-normal demand

As stated in Chapter 1, the aim of this research is to establish an appropriate forecasting strategy for predicting the demand for spare parts in the South Korean Navy. Much of the research in forecasting non-normal demand has only been carried out in terms of a direct forecasting method: exponential smoothing (Croston, 1972, Sani and Kingsman, 1997, Narasimhan et al., 1998, Ghobbar and Friend, 2003), Croston's method (Croston, 1972, Willemain et al., 1994, Johnston and Boylan, 1996, Sani and Kingsman, 1997), Syntetos-Boylan approximation (Syntetos and Boylan, 2001, Eaves, 2002, Eaves and Kingsman, 2004, Syntetos and Boylan, 2005), weighted moving average (Ghobbar and

Friend, 2003, Regattieri et al., 2005), and the Box and Jenkins models (Businger and Read, 1999).

Hierarchical forecasting was discussed in many application areas such as economics, marketing, manufacturing, planning and travelling (Theil, 1954, Shlifer and Wolff, 1979, Wei and Abraham, 1981, Schwarzkopf et al., 1988, Fliedner and Lawrence, 1995, DeLurgio, 1998, Kahn, 1998, Fliedner, 1999, Fliedner, 2001, Widiarta et al., 2006, Hyndman et al., 2007, Viswanathan et al., 2008, Widiarta et al., 2008a, Widiarta et al., 2009). However, the nature of spare parts demand might be rather different. The spare parts demand is more intermittent and more variable (Ghobbar and Friend, 2002). It usually comprises of many periods of no demand and its demand size is highly variable.

The applicability of hierarchical forecasting to predicting spare parts demand in the South Korean Navy was stated in Section 1.5. However, the literature has paid little attention to the use of hierarchical forecasting for the intermittent demand at item level. This is a feature of non-normal demand associated with spare parts demand. As stated above, Fliedner and Mabert (1992) and Fliedner and Lawrence (1995) examined the performance of hierarchical forecasting with the automotive spare parts demand. However, they screened out irregular time series such as time series with missing observations and demand values of zero. Although they found the volatile demand feature which is a demand feature of non-normal demand, the intermittent demand feature was not examined.

Recently, Viswanathan et al. (2008) applied hierarchical forecasting to intermittent group level demand in a simulation study. The intermittent demand was generated by a variety of probability distributions. However, their discussion was restricted to forecasting at group level and a production planning context with data generated from probability distributions. Forecasting intermittent demand at item level with empirical data might be different.

In this section, the comparative performance of forecasting strategies was reviewed. Much of the research has found that the performance of hierarchical forecasting is conditional on some specific demand features. Lack of agreement in the performance of hierarchical forecasting strategies, which might originate from the different features of the data, was identified. Lack of hierarchical forecasting research for intermittent demand at item level with empirical data was also identified. Therefore, a literature review about the effect of demand features including the intermittent demand feature upon the performance of direct and hierarchical forecasting strategies is necessary.

2.5 Classification of Demand for Forecasting

This section reviews the literature about the classification scheme of demand and the influence of demand features upon the performance of forecasting methods. Classification schemes for non-normal demand and the influence of non-normal demand features upon the performance of direct forecasting are reviewed. This is followed by a literature review about the influence of demand features upon the performance of hierarchical forecasting. Then, the literature about the influence of demand features upon the performance of hierarchical forecasting is evaluated with respect to research opportunities.

2.5.1 Direct forecasting

Various demand features have been considered for classifying non-normal demand for direct forecasting (Williams, 1984, Willemain et al., 1994, Businger and Read, 1999, Ghobbar and Friend, 2002, Ghobbar and Friend, 2003, Eaves and Kingsman, 2004, Regattieri et al., 2005, Syntetos, 2007, Boylan et al., 2008, Eaves, 2002). Their classification schemes are reviewed below.

Businger and Read (1999) classified the quarterly demand time series for spare parts in the US Navy in order to predict the suitability of the Box-Jenkins models to each class of the demand time series. In the first step, they classified the demand time series for spare parts by two dimensions such as the coefficient of variation in demand size and the number of periods with zero demand.

Coefficient of variation in demand size is a unit free measure of relative variability (Williams, 1984, Businger and Read, 1999, Syntetos, 2001). The erratic and the lumpy demand features in the categories of non-normal demand (Boylan et al., 2008, p. 474) might be captured by this statistic. It can be expressed as:

$$\textit{Coefficient of variation in demand size} = s/\bar{y} \qquad (2\text{-}32)$$

where: s = the standard deviation of demand size

$\bar{y}$ = the mean demand size

The number of periods with zero demand was used in order to measure the intermittency of a demand time series (Businger and Read, 1999, Boylan et al., 2008). Demand features reflecting intermittency (i.e. intermittent, slow moving, clumped and lumpy) in the categories of non-normal demand (Boylan et al., 2008) might be captured by this statistic. However, this statistic cannot provide a general measure for classification. This statistic depends on the overall data periods; that is, the number of zero periods might be longer when the overall data periods are longer, and vice versa.

Table 2-3 Classification of spare parts demand by two dimensions (Businger and Read, 1999)

Zero count cell group	Coefficients of variation groups				
	$0 \le s/\bar{y} < 0.8$	$0.8 \le s/\bar{y} < 1.0$	$1.0 \le s/\bar{y} < 1.2$	$1.2 \le s/\bar{y} < 1.5$	$1.5 \le s/\bar{y}$
1	0	0-2	0-5	0-8	0-8
2	-	3-5	6-9	9-12	9-12
3	1	6-7	10-12	13-16	13-16
4	2-3	8-10	13-15	17-18	17-18
5	4-20	11-20	16-20	19-20	19-20

Businger and Read (1999) partitioned the coefficient of variation in demand size into five groups by the four boundaries (i.e. 0.8, 1.0, 1.2 and 1.5); then, each group was re-partitioned into five groups by the number of periods with zero demand as shown in Table 2-3. Roughly the same number of the demand time series for spare parts was allocated to each cell. The numbers in each cell indicate the boundaries of the number of zero demand series allocated in each cell. Businger and Read (1999) claimed that each cell has some internal homogeneity in terms of the coefficients of variation in demand size and the number of periods with zero demand. In the second step, although the variations and the intermittency of demand can be considered to be particular characteristics of non-normal demand, they considered more statistics such as trend, the number of peaks, skewness, and autocorrelation.

Trend was computed as in equation (2-33) (Businger and Read, 1999). The time series are divided into thirds. A trend is not less than -1 and not more than +1, because y_L is not less than $y_{(1/6)}$, and y_U is not more than $y_{(5/6)}$.

$$Trend = (y_U - y_L)/(y_{(5/6)} - y_{(1/6)}) \tag{2-33}$$

where: y_U = the upper third median

y_L = the lower third median

$y_{(1/6)}$ = the 1/6 median

$y_{(5/6)}$ = the 5/6 median

The number of peaks was calculated as in equation (2-34) (Businger and Read, 1999). $I(d_t > 2)$ is an indicator function taking the value of 0 if $d_t \leq 2$ or 1 otherwise. $d_t \leq 2$ indicates that y_t is a non-peak demand; $d_t > 2$ indicates that y_t is a peak demand. The erratic and the lumpy demand features (Boylan et al., 2008) might be captured by this statistic.

$$Number\ of\ peaks = \sum_{t=1}^{n} I(d_t > 2) \tag{2-34}$$

where: y_t = the observed demand size for an item at time period t

$\bar{y}$ = the mean demand size

s = the standard deviation of demand size

n = the number of the total periods

$$d_t = \frac{|y_t - \bar{y}|}{s}$$

$$I(d_t > 2) = \begin{cases} 0, & d_t \leq 2 \\ 1, & d_t > 2 \end{cases}$$

The lack of symmetry of a distribution around its mean can be identified by skewness

(Tabachnick and Fidell, 2007). The values, above or below zero, indicate departure from a normal distribution (Miles and Shevlin, 2001, Tabachnick and Fidell, 2007, Howitt and Cramer, 2008). A simple index of skewness was measured as in equation (2-35) (Businger and Read, 1999).

$$Skewness = \bar{y} / m \tag{2-35}$$

where: $\bar{y}$ = the mean demand size; m = the median demand size

Autocorrelation is defined as "the correlation between values of the same time series at different time periods" (Hyndman et al., 1998, p. 590). Autocorrelation (Willemain et al., 1994, Businger and Read, 1999) and cross-correlation (Willemain et al., 1994) were examined to identify demand features for non-normal demand.

Businger and Read (1999) examined the relationships between the accuracy of the four models including three Box-Jenkins models [ARIMA(1, 1, 1), ARIMA(2, 2, 2) and ARIMA(3, 2, 3)] and exponential smoothing and the spare parts demand time series characterised by the above statistics. They used multiple linear regressions to analyse the relationships. The independent variables in the regressions were the above statistics; the dependent variable was an objective function, namely measure of effectiveness (MOE), as expressed in equation (2-36). The greater MOE score for a forecasting method indicates the greater accuracy of the forecasting method.

$$MOE = s / \sqrt{\sum (y_t - \hat{y}_t)^2 / (n-1)} \tag{2-36}$$

where: s = the standard deviation; $\hat{y}_t$ = the fitted value; and n = the number of the total periods

Businger and Read (1999) found that the Box-Jenkins models are superior to exponential smoothing and ARIMA (1, 1, 1) is the best. However, they failed to find distinct patterns in the relationships between the forecasting accuracy and the statistics.

Williams (1984) and Eaves and Kingsman (2004) considered lead time variation in the demand classification. Williams (1984) established an analytical method for classifying demand for 11,000 products over two years, which are generally low volume. Each product was regarded as independent and the demand was regarded to be entirely unplanned. He decomposed the variance of the demand during a lead time (DDLT) into the three constituent parts of transaction variability ($C_n^2/\overline{L}$), demand size variability ($C_x^2/\overline{n}\overline{L}$) and lead-time variability (C_L^2). The variance partition equation can be written as in equation (2-37).

$$C_{DDLT}^2 = \frac{C_n^2}{\overline{L}} + \frac{C_x^2}{\overline{n}\overline{L}} + C_L^2 \qquad (2\text{-}37)$$

where:

C_{DDLT}^2 = the square coefficient of variation of DDLT

C_n^2 = the square coefficient of variation of number of orders, n

C_x^2 = the square coefficient of variation of order sizes, x

C_L^2 = the square coefficient of variation of lead times, L

$\overline{L}$ = the mean replenishment lead time

$\overline{n}$ = the mean number of orders per unit time

Williams (1984) then developed a two-dimensional classification scheme (i.e. the transaction variability and the demand size variability) in order to clarify demand features of the products for applying a suitable forecasting method to each product. If the number of demands for an item per week is Poisson distributed with mean λ, with a constant lead time, the transaction variability can be represented as $1/\lambda\bar{L}$ (boundary = 0.7), and the demand size variability (i.e. lumpiness of the demand) can be represented as $C_x^2/\lambda\bar{L}$ (boundary = 0.5). This classification scheme can be expressed as Figure 2-1. While he argued that the choice of boundaries between each of the categories is a management decision, he offered no rational criterion to set these boundaries.

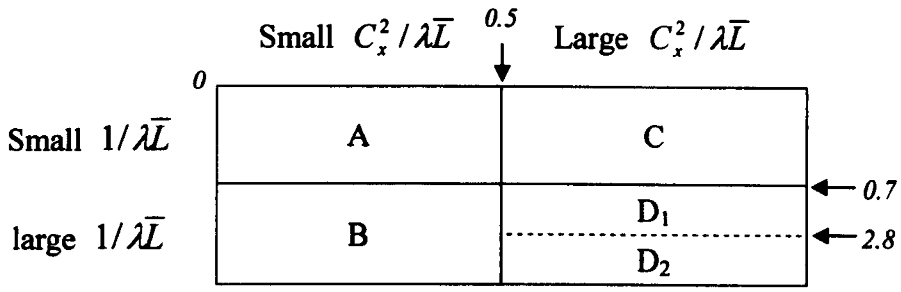


Figure 2-1 Two-dimensional classification of demand during lead time (Williams, 1984)

Williams (1984, p. 939) defined sporadicity as: "high sporadicity - one demand at least ten times the average weekly demand; low sporadicity – average demand during the lead time less than ten items; and no sporadicity – neither of the above". Then, he argued that class A mainly has rapidly arriving demand of a not-widely spreading demand size (namely no sporadicity); class B mostly has slow-moving demand (namely low-sporadicity); class C mostly has frequent demand of widely varying demand sizes; and class D mostly has little demand of widely varying demand sizes (namely high sporadicity). He sub-classified class D into D_1 and D_2 using a boundary of 2.8 for $\lambda\bar{L}$: $0.7 \leq D_1 < 2.8$; and $2.8 \leq D_2$. Then, he argued that Croston's method has to be used for class C and D_1. Although Williams (1984) presented a sophisticated classification scheme, he offered no rational criterion to set those boundaries. He also made no attempt to provide an empirical evidence for his argument on forecasting methods.

Eaves and Kingsman (2004) crystalised the analytical framework of Williams (1984). They categorised the time series of spare parts demand in the UK Air Force into smooth, irregular, slow-moving, and intermittent demand features. Then, they distributed the demand features by transaction variability, demand size variability, and lead-time variability as shown in Table 2-4.

Table 2-4 Classification of demand (Eaves and Kingsman, 2004)

Lead-time demand component			Demand pattern classification
Transaction variability	Demand size variability	Lead-time Variability	
Low	Low		Smooth
Low	High		Irregular
High	Low		Slow moving
High	High	Low	Mildly intermittent
High	High	High	Highly intermittent

In common with Williams (1984), Eaves and Kingsman (2004) argued that the choice of boundaries between each of the categories is a management decision. The boundaries of demand size variability and lead-time variability were set to the median of the demand size variability and the median of the lead-time variability respectively. However, the boundary of transaction variability was decided at a lower quartile of the transaction variability because of the higher variability of the spare parts in the UK Air Force compared to most other organisations. Thus, 25% of the demand was classified as smooth or irregular, approximately half the remainder was classified as slow moving, and the remainder was classified as intermittent (either mildly or highly intermittent).

As stated earlier, Eaves and Kingsman (2004) then examined the performance of

Syntetos-Boylan approximation (Syntetos and Boylan, 2005), exponential smoothing, moving average and Croston's method for the classification using an inventory model simulation. They claimed that Syntetos-Boylan approximation minimises stock-holdings across all the demand categories.

The above mentioned classification schemes appear to be built in a sophisticated fashion. However, the above research failed to present a clear guideline for selecting a forecasting method suitable to forecast a demand in a category. There is research (Johnston, 1980, Johnston and Boylan, 1996, Boylan et al., 2008) which has presented a clear guideline for selecting a forecasting method.

Average inter-demand interval (ADI) refers to the average period of time in which a demand occurs. Average inter-demand interval was used in order to measure the intermittency of a demand (Johnston, 1980, Johnston and Boylan, 1996, Syntetos, 2001, Boylan et al., 2008). Some of the demand features (i.e. intermittent, slow moving, clumped and lumpy) in the categories of non-normal demand (Boylan et al., 2008, p. 474) might be captured by this statistic.

As stated earlier, Johnston and Boylan (1996) suggested a guideline for Croston's method and exponentially weighted moving average as a result of simulation analysis with generated data. Johnston and Boylan (1996) argued that if the average inter-demand interval is greater than 1.25 forecast review periods (i.e. weeks), then Croston's method should be used rather than exponentially weighted moving average; whereas the variability of demand sizes is considered to have no effect on forecasting performance.

Boylan et al. (2008) developed the guideline and suggested a classification scheme that allocates forecasting methods according to the number of periods with zero demand and average inter-demand interval during the last *n* time periods as shown in Table 2-5. The last 13 time periods (n = 13) for the classification was used. However, they offered no explanation for the decision of *n*.

Table 2-5 Classification scheme for identifying intermittence (Boylan et al., 2008)

	Intermittent	Non-intermittent	Recommended range of break-point values (number of zeroes)	Equivalent range of average inter-demand intervals
1	Croston	SMA	5 – 7	1.63 – 2.17
2	Croston	SES	6 – 8	1.86 – 2.60
3	SBA	SMA	2 – 3	1.18 – 1.30
4	SBA	SES	2 – 4	1.18 – 1.44

Boylan et al. (2008) compared each pair of forecasting methods in terms of forecasting accuracy for 16,000 empirical data such as monthly automotive spare parts data over 26 periods, bi-monthly aerospace spare parts data over 52 periods, and monthly chemical products data over 60 periods. They considered Croston's method (Croston, 1972) and the Syntetos-Boylan approximation (SBA) (Syntetos and Boylan, 2005) to be more suitable for the intermittent demand category (i.e. demand features associated with a 'low' demand frequency as determined by the break-point), whereas simple exponential smoothing (SES) and simple moving average (SMA) were considered to be more suitable for non-intermittent demand category (i.e. demand features associated with a 'high' demand frequency). For example, when nine zero demand periods are observed in a data set during the last 13 periods, Croston's method is preferred to simple moving average 13 (i.e. 13 periods) and simple exponential smoothing.

In this subsection, various classification schemes for non-normal demand and various attempts to guide the selection of direct forecasting methods using a classification scheme were reviewed. Some limitations of the research were also identified. In the next subsection, a literature review about the classification schemes and the guidelines for the use of hierarchical forecasting is presented.

2.5.2 Hierarchical forecasting

As stated earlier, most research has demonstrated that the performance of top-down forecasting is conditional on some specific demand features. These demand features might be able to guide the selection of a forecasting strategy for data characterising a specific demand feature. Table 2-6 compares the results of major studies that have compared the influences of demand features upon the relative performance of top-down and direct forecasting strategies. Superior forecasting strategies at group or item level, which are conditional on the value (high or low) of the demand features in the 1st column, are presented in the 2nd or 3rd column respectively. These are reviewed as follows.

Table 2-6 The influence of demand features upon the relative performance of top-down forecasting and direct forecasting

Demand feature	Impact on relative performance [superior strategy (level)] High	Impact on relative performance [superior strategy (level)] Low	Reference
Correlation	Positive: DF (item) Negative: TD	TD	Schwarzkopf et al (1988)
		DF (item)	Gross and Sohl (1990) & Widiarta et al.(2006)
	Both DF (group) and BU		Fliedner (1999)
	Non-significance (DF outperforms TD)		Dangerfield and Morris (1992)
	Non-significance (identical performance)		Widiarta et al. (2008a) & Widiarta et al. (2009)
Proportion	Greater superiority in DF (item)	DF (item)	Dangerfield and Morris (1992)
	Non-significance		Widiarta et al.(2006)
	Non-significance (identical performance)		Widiarta et al.(2008a)
Variability	TD	DF (item)	Schwarzkopf et al. (1988)
	DF (group)	BU	Viswanathan et al.(2008)
Substitutability (β) & variability of proportion (υ)	High β (& high υ) or Low υ: TD Low~medium β & High υ: DF (item)		Widiarta et al. (2008b)
Forecasting horizon	Long: TD	Short: DF (item)	Shlifer and Wolff (1979)
Lag-1 auto correlation [$\rho(1)$]	$\rho(1) > 1/3$: DF (item)	$-1 < \rho(1) < 1/3$: Non- significance	Widiarta et al.(2006)
Grouping criteria	UV and DV are better than SI and HP		Fliedner and Mabert (1992)
No. of groups (No. of items in a group)	Non-significance		Fliedner and Mabert (1992) Fliedner and Lawrence (1995)

TD = top-down forecasting; BU = bottom-up forecasting; DF (group or item) = direct forecasting (at group or item level); UV= historical unit volume; DV = historical dollar volume; SI = item demand series' seasonal index; HP = historical item forecasting performance.

Correlations

Previous research findings into the influence of correlations between item level time series in a group upon the relative performance of top-down and direct forecasting strategies have been inconsistent and contradictory as shown. Schwarzkopf et al. (1988) compared top-down forecasting and direct forecasting using an analytic method. Two item level top-down forecasts of demand i (i = 1, 2), τ periods ahead made at time t, $f_{i,t+\tau}$ can be expressed as $f_{1,t+\tau} = PR_{1,t+\tau}(\widetilde{F}_{t+\tau} + \varepsilon_{1,t+\tau})$ and $f_{2,t+\tau} = PR_{2,t+\tau}(\widetilde{F}_{t+\tau} + \varepsilon_{2,t+\tau})$ where $\widetilde{F}_{t+\tau}$ is a group level direct forecast composed of the two items, τ periods ahead made at time t; $PR_{i,t+\tau}$ is the estimated percentage for item i ($PR_{1,t+\tau}+ PR_{2,t+\tau} = 1$); and $\varepsilon_{i,t+\tau}$ is the estimated error terms assumed to be independent with mean 0 and variance s^2.

Schwarzkopf et al. (1988) contended that, when two item level time series are independent, the sum of the variability of top-down forecasting can be expressed as in equation (2-38). Letting $PR_{1,t+\tau} = PR_{2,t+\tau} = 1/2$, the sum of the variability of the errors of top-down forecasting is = $s^2/4$; and the sum of the variability of the errors of direct forecasting is = $s^2/2$ (i.e. the errors of direct forecasting are more variable than the errors of top-down forecasting). However, if the time series of items are dependent, which is usually the case, a covariance term, $Cov(f_{1,t+\tau}, f_{2,t+\tau})$, has to be incorporated into the relationships for top-down forecasting as in equation (2-39). Thus, the sum of the variability of the errors of top-down forecasting is greater than the sum of the variability of the errors of direct forecasting when there is a strong positive correlation between the items. However, the variability of the errors of top-down forecasting is smaller when there is a negative correlation.

$$Var(f_{1,t+\tau} + f_{2,t+\tau}) = Var(f_{1,t+\tau}) + Var(f_{2,t+\tau}) \tag{2-38}$$

$$Var(f_{1,t+\tau} + f_{2,t+\tau}) = Var(f_{1,t+\tau}) + Var(f_{2,t+\tau}) + 2Cov(f_{1,t+\tau}, f_{2,t+\tau}) \tag{2-39}$$

As stated earlier, Gross and Sohl (1990) observed very low correlations among all the time series for their analysis both within and between the product lines. They therefore suggested that direct forecasting outperformed top-down forecasting consistently for these time series. Widiarta et al. (2006) conceded to the argument of Gross and Sohl (1990) about correlations. Widiarta et al. (2006) examined the relative performance of top-down forecasting and direct forecasting with AR (1) process. Exponential smoothing was used as the forecasting method for both forecasting strategies. Equation (2-22) was used as the proration method for top-down forecasting. They contended that the supremacy of direct forecasting increases when the correlation between two items decreases.

These arguments (Gross and Sohl, 1990, Widiarta et al., 2006) with the argument of Fliedner (1999) about correlations seem to be inconsistent with the above arguments of Schwarzkopf et al. (1988). While the latter claimed the superiority of top-down forecasting with a decreasing correlation, the former claimed either the superiority of direct forecasting with a decreasing correlation (Gross and Sohl, 1990, Widiarta et al., 2006) or the superiority of forecasts at group level with an increasing correlation (Fliedner, 1999). The argument (Fliedner, 1999) is considered as being in line with the arguments of Gross and Sohl (1990) and Widiarta et al. (2006) because the superiority of forecasts at group level implies an applicability of top-down forecasting at item level (Gross and Sohl, 1990).

A number of authors have also found that a correlation does not have any significant effect upon the relative forecasting performance between direct forecasting and top-down forecasting (Dangerfield and Morris, 1992, Widiarta et al., 2008a, Widiarta et al., 2009).

Proportion

There is research which has examined the influence of an item's proportion of the group, in which the item is a member, upon the relative performance of top-down forecasting and direct forecasting. An item's proportion was argued not to have significant influence upon the relative performance of top-down forecasting and direct forecasting in previous research. Dangerfield and Morris (1992) investigated an item's proportion of a group as well as the effects of correlations between items in a group upon forecasting performance. As stated above, they found that direct forecasting is superior to top-down forecasting in almost three out of four data sets, regardless of the items' proportions. Greater superiority of direct forecasting was observed when an item's proportion (p_i) of the group was high ($0.65 < p_i < 1.0$). This non-significance of the proportion for the relative performance was conceded to by the two pieces of research (Widiarta et al., 2006, Widiarta et al., 2008a).

Variability

Schwarzkopf et al. (1988) argued that direct forecasting is to be preferred in order to detect distinctions between demand patterns for individual items. However, when individual demand data contains missing or unreliable data, then top-down forecasting is better.

As stated in Subsection 2.4.1, Viswanathan et al. (2008) indicated that forecasting performance at group level depends on the variability of demand interval and demand size; that is, when the variability of interval is low, bottom-up forecasting is superior; and conversely, when the variability of interval and demand size is high, direct forecasting is superior. This might suggest the applicability of top-down forecasting to a time series which is highly variable in the demand interval and the demand size. This is because the superiority of forecasts at group level implies an applicability of top-down forecasting at item level (Gross and Sohl, 1990).

Although the demand interval and the demand size (Viswanathan et al., 2008) are completely different demand features from the missing or unreliable data (Schwarzkopf et al., 1988), the argument of Schwarzkopf et al. (1988) might be more or less in line with Viswanathan et al. (2008) in that both arguments can be interpreted as the applicability of top-down forecasting for highly variable data.

Substitutability and variability of proportion

Widiarta et al. (2008b) claimed that the degree of substitutability and the variability of an item's proportion influenced the relative performance of top-down forecasting and direct forecasting after a simulation experiment with three time series [i.e. MA (1), AR (1) and ARMA (1, 1)]. Exponential smoothing was used as a direct forecasting method for both top-down and direct forecasting strategies. Equation (2-26) was used as the proration method for top-down forecasting. The number of items in a group was restricted to two items. The degree of substitutability, β_{ij}, indicates a portion of the unsatisfied demand for item *i* that is passed to item *j*, on condition that item *j* has excess inventory. The variability of an item's proportion in the group, υ, indicates the range of variability of the item's proportion in the group. υ was calculated as the upper bound (*UB*) of the item demand in the group minus the lower bound (*LB*) of the item demand in the group ($\upsilon = UB - LB$). The demand proportion for each item in the group, p_i ($0 \leq p_i \leq 1$), was assumed to be uniformly distributed, p_i U (LB, UB).

Widiarta et al. (2008b) argued that, at group level, direct forecasting outperformed bottom-up forecasting consistently. The relative advantage of direct forecasting was claimed to be more augmented as the degree of substitutability, β, decreased and the variability of the product's proportion, υ, increased. At item level, the superiority of top-down forecasting was argued to increase when the degree of substitutability, β, increases (approximately greater than 0.2).

Widiarta et al. (2008b) suggested that the reason for this result was as follows. When two items are highly substitutable, the information distortion of the observed demand from real demand becomes more pronounced. Any excess demand, β, for an item A would be satisfied by a substitutable item B, and this phenomenon is invisible to the inventory manager. Thus, the inventory manager is prone to under-forecasting the demand of item A, but over-forecasting the demand of item B. However, top-down forecasting is less affected by the information distortion because top-down forecasting uses the historical demand of the items' group, which could be less dependent upon the degree of item substitutability, β. Assuming that substitutability of Naval spare parts demand is high, top-down forecasting might be superior to direct forecasting at item level.

Widiarta et al. (2008b) continued their argument that top-down forecasting outperformed direct forecasting with the low variability of proportion. However, as the variability of proportion increased, direct forecasting outperformed top-down forecasting with the low to medium substitutability where the low to medium substitutability is a value between 0.0 and 0.8 for MA (1), between 0.0 and 0.2 for AR (1), and 0.0 for ARMA (1, 1). They suggested that the reason for this result was that, as the variability of an item's proportion increases, the amplification of the item's demand increases, so that this makes the proration process of top-down forecasting more difficult.

Forecasting horizon

As stated earlier, Shlifer and Wolff (1979) contended that direct forecasting is preferred to top-down forecasting. Then, they argued that for some time boundary, θ, direct forecasting is superior for forecasting horizon, $t < \theta$; however, top-down forecasting is superior for $t > \theta$. As a practical implication of this arithmetic induction, they contended that, direct forecasting is preferred to top-down forecasting for the near future; however, top-down forecasting is preferred as the forecast goes further into the future. It implies that, when a forecasting horizon is long, this might be an advantageous condition for top-down forecasting.

Lag-1 autocorrelation

Widiarta et al. (2006) postulated that the forecasting performance depends on lag-1 autocorrelation of a demand time series. When the lag-1 autocorrelation, $\rho(1)$, of the time series for at least one of the items in a group consisting of two items was greater than 1/3, direct forecasting outperformed top-down forecasting. However, when the lag-1 autocorrelations of the two item level time series were satisfied, $-1 < \rho(1) \leq 1/3$, the difference in the performance of the two forecasting strategies was non-significant.

Grouping criteria

The relationships between grouping (or clustering) criteria and the performance of top-down forecasting have been investigated. Fliedner and Mabert (1992) examined grouping criteria and the influence of the number of groups (i.e. the number of items in a group) upon the performance of top-down forecasting using empirical product data (i.e. demand for automotive spare parts from Cummins Engine Inc.). As stated earlier, Cummins Engine Inc. uses a hierarchical structure for classifying products; namely standard product classification. In total, 22,241 monthly demand data for spare parts such as cylinder heads, turbocharger components, fuel pump components and air compressors, were used for their study. Similarly to Fliedner and Lawrence (1995), Fliedner and Mabert (1992) screened out time series with missing observations and demand values of zero and found the volatile demand feature that might be a demand feature of non-normal demand. Holt-Winters forecasting was used for top-down forecasting; equation (2-23) was used for the proration method.

Fliedner and Mabert (1992) analysed the performance of top-down forecasting with respect to: a) four grouping criteria (UV: historical unit volume; DV: historical dollar volume; SI: item demand series' seasonal index; and HP: historical item forecasting performance); and b) the number of groups (i.e. two, five and ten) with 22,241 item level time series. Items that have similar volumes of historical demand formed a group using UV grouping criterion. Similar items based on the product of historical demand volumes and item dollar values formed a group using DV grouping criterion. An analysis of variance (ANOVA) identified that grouping based on UV and DV provided significantly smaller mean absolute percentage error (MAPE) and mean percentage error (MPE) for top-down forecasting than grouping based on SI and HP. Analysis of variance (ANOVA) refers to a statistical method for comparing two or more groups in terms of their mean scores on a dependent variable (Howitt and Cramer, 2008).

Variations in the number of groups with a limited number of items indicates variations in the number of items within a group (Fliedner and Lawrence, 1995). The number of items within a group (or the number of groups with the limited number of items) was not observed to provide a significant difference in the performance of top-down forecasting (Fliedner and Mabert, 1992). Fliedner and Lawrence (1995) corroborated that an ANOVA identified the non-significant difference of the number of groups with the limited number of items (or the number of items within a group) in the performance of top-down forecasting.

Seasonality in hierarchical forecasting

Kahn (1998) briefly suggested guidelines for a hierarchical forecasting strategy in terms of seasonality in his empirical study with sales data consisting of three levels (top level: one brand; intermediate level: two items; and bottom level: seven locations). He suggested that a combinatorial forecasting method, which combined a non-seasonal top level direct forecast with a seasonal bottom level direct forecast, provided superior forecasting at the bottom level.

Dekker et al. (2004) argued that Holt-Winters forecasting, which can handle trend and seasonality, performed worse than simple exponential smoothing. They reasoned that the seasons in the data are stochastic (e.g. the summer may begin later one year and earlier the next year). Then, as mentioned in Subsection 2.4.3, they argued that the aggregation method combined with Naive 1 (i.e. a forecast identical with the most recent demand) could minimise forecasting errors.

Direct forecasting method within hierarchical forecasting strategy

As for a direct forecasting method within hierarchical forecasting strategy, simple exponential smoothing has been evaluated as superior to moving average or other complex models such as Holt-Winters forecasting and ARIMA models in several studies (Fliedner and Lawrence, 1995, Fliedner, 1999, Dekker et al., 2004, Viswanathan et al., 2008). Fliedner (1999) recommended simple exponential smoothing for hierarchical forecasting strategy owing to the following reasons:

a) it is frequently selected for industrial production planning and inventory control purposes;
b) it requires less time to generate forecasts especially with a large number of items; and
c) it is suitable for demand without such patterns as seasonality and trend.

The fit of the exponential smoothing parameter to data was contended not to have a significant effect on the relative performance of top-down forecasting and direct forecasting (Dangerfield and Morris, 1992). In their experiment, Dangerfield and Morris (1992) compared the best fit smoothing parameters and randomly selected smoothing parameters, and found that the selection of a smoothing parameter was not a substantial factor of the relative forecasting performance.

2.5.3 Evaluation of hierarchical forecasting research

Previous research findings into the influence of demand features upon the performance of hierarchical forecasting have been rather inconsistent and contradictory as shown in Table 2-6. There has also been little discussion about the influence of non-normal demand features upon the relative performance of hierarchical and direct forecasting methods. Thus, several research opportunities can be identified in these areas. This subsection discusses these controversial issues and research gaps.

Disagreement in the influence of correlations upon hierarchical forecasting

The influence of correlations upon the relative performance of top-down forecasting and direct forecasting might be the most controversial issue (Schwarzkopf et al., 1988, Gross and Sohl, 1990, Dangerfield and Morris, 1992, Fliedner, 1999, Widiarta et al., 2006, Widiarta et al., 2008a). As stated above, the disagreement could be attributed to different features of data in each investigation (Kahn, 1998, Fliedner, 1999).

In much of the research about the influence of correlations upon the performance of top-down forecasting and direct forecasting, the number of items in a group was restricted to two items. However, the statistical features of the data used in the research were very different. This is shown in Table 2-7. Much of the research for correlations used simulated data rather than empirical data. Some studies (Fliedner, 1999, Widiarta et al., 2009) were about group level forecasting. These might be the main sources of the disagreement.

Table 2-7 Research on the influence of correlations upon hierarchical forecasting

Reference	Impact on relative performance	Forecasting level	No. of items in a group	Data
Schwarzkopf et al. (1988)	↓TD	Item	2	Analytic study
Gross and Sohl (1990)	↓DF(item)	Item	3 ~ 7	Steel sales
Widiarta et al. (2006)		Item	2	AR (1)
Fliedner (1999)	↑DF(group) & BU	Group	2	MA (1)
Dangerfield and Morris (1992)	Non-significance	Item	2	M-competition
Widiarta et al. (2008a)		Item	2	MA (1)
Widiarta et al. (2009)		Group	2	MA (1)

↑ (or ↓) = increasing (or decreasing) correlations increases the relative performance of the forecasting strategy; TD = top-down forecasting; DF(group or item) = direct forecasting at group or item level; BU = bottom-up forecasting.

As stated above, the nature of non-normal demand is rather different from the demand used in the research on the impact of correlations on the performance of top-down forecasting. However, there have been no controlled studies which examine the influence of correlations between non-normal demand time series upon the relative performance of hierarchical and direct forecasting methods.

The influence of intermittency upon hierarchical forecasting

As shown in the categories of non-normal demand, high variability and intermittency seem to be important demand features for non-normal demand. There is research (Schwarzkopf et al., 1988, Viswanathan et al., 2008) which can be interpreted as arguing for the applicability of top-down forecasting for highly variable data.

Intermittency is a demand feature that makes forecasting very difficult (Willemain et al., 1994, Syntetos and Boylan, 2005). As stated earlier, the intermittency is an important demand feature for the selection of a direct forecasting method. The influence of intermittency (e.g. number of periods with zero demand) and variance (e.g. coefficient of variation in demand size) upon the performance of direct forecasting methods has been well-identified. This is especially so with Johnston and Boylan (1996) and Boylan et al. (2008) who suggested clear cut-off values to classify intermittent demand for direct forecasting methods. However, no research has been carried out on the influence of intermittency upon the performance of hierarchical forecasting.

Influence of demand features upon combinatorial forecasting

All the literature regarding the influence of demand features upon hierarchical forecasting was carried out for top-down forecasting. In practice combinatorial forecasting was contended to be a superior forecasting strategy to top-down forecasting (Kahn, 1998, Dekker et al., 2004, Hyndman et al., 2007). What is needed is an investigation on the influence of demand features upon the performance of combinatorial forecasting. However, there may be no controlled research which examines the influence of demand features upon the performance of combinatorial forecasting.

Combined influence of demand features upon hierarchical forecasting

Non-normal demand encompasses various demand features as shown in Subsection 1.2.2. In order to capture the nature of non-normal demand for selecting a superior forecasting method, a multidimensional calibration of data features might be necessary.

Some researchers classified non-normal demand for direct forecasting methods in terms of multiple demand features (Williams, 1984, Businger and Read, 1999, Eaves and Kingsman, 2004). However, they failed to find any empirical evidence for a distinct combined influence of demand features upon direct forecasting methods. Williams (1984) provided no empirical evidence; the Box-Jenkins models have been observed to be superior regardless of the demand features (Businger and Read, 1999); and Syntetos-Boylan approximation has also been observed to be superior regardless of the demand features (Eaves and Kingsman, 2004).

Some researchers also examined the influence of multiple demand features upon the performance of top-down forecasting (Schwarzkopf et al., 1988, Dangerfield and Morris, 1992, Widiarta et al., 2006, Widiarta et al., 2008a). However, they have failed to find any combined influence of demand features upon the relative performance between top-down forecasting and direct forecasting. Schwarzkopf et al. (1988) examined multiple demand features such as correlations and reliability of demand; however, they did not take into account the combined influence of demand features. Dangerfield and Morris (1992) examined a combined influence of the correlation and item's proportion in the group upon the relative performance. However, no combination of the correlation and item's proportion that influences on the relative performance was found. Widiarta et al. (2006) investigated a combined influence of lag-1 autocorrelation, an item's proportion in the group and the correlation. However, they found that the relative forecasting performance mainly depends on the lag-1 autocorrelation regardless of the item's proportion and the correlation. Widiarta et al. (2008a) argued that the correlation between the demand time series of items and the item's proportion in the group were non-significant demand features for the relative forecasting performance.

Widiarta et al. (2008b) identified the combined influence of the degree of substitutability and the variability of an item's proportion upon the relative performance of top-down forecasting and direct forecasting. Their argument was restricted to a simulation experiment with three time series [i.e. MA (1), AR (1) and ARMA (1, 1)]. The relationships between the features of non-normal demand and the relative performance of hierarchical forecasting might be different. However, they provided no evidence for their findings with empirical non-normal time series.

In this section much of the previous research on classification schemes for non-normal demand for the selection of direct forecasting methods and the influence of demand features upon the relative performance of hierarchical forecasting were reviewed. Some limitations of the research, some controversial issues, and several research opportunities were also identified.

In the literature reviewed above, various accuracy measures were used with their own justification; however, those measures produced differing results. For instance, Dangerfield and Morris (1992) found that the results from mean absolute percentage error (MAPE) and mean squared error (MSE) were very different. While top-down forecasting was observed to be superior to direct forecasting for 51 ~ 63% of the cases in terms of MSE, direct forecasting was observed to be superior to top-down forecasting for 73 ~ 74% of the cases in terms of MAPE. Then, they used the results from MAPE because they believed that the results from MSE were unreliable. Establishing reliable accuracy measures is important. This requires a discussion about forecasting accuracy measurements.

2.6 Measures of Forecasting Accuracy

Measuring forecasting performance is a crucial issue. This is because different accuracy measures can lead to different conclusions (Syntetos and Boylan, 2005). The best model under one criterion cannot always be the best under some other criteria (Chatfield, 2004). A performance measure also contributes to calibrating or refining a model in order to forecast accurately in a given set of time series (Armstrong and Collopy, 1992). Moreover, the special property of intermittent data (i.e. some zero-demand periods) creates a particular difficulty in selecting an appropriate accuracy measure (Syntetos and Boylan, 2005). Various accuracy measures are available. Syntetos and Boylan (2005) divided accuracy measures into two groups, namely absolute measures and relative measures: a) an absolute measure of error evaluates a forecasting method in isolation; and b) a relative measure evaluates one forecasting method relative to another method across a set of time series, where the forecast error of the each method is evaluated using one of the absolute measures. A limitation of these measures is that they do not capture the monetary value and the service level of each item, so do not measure the practical impact that a forecasting method has on the inventory system. A derivative measure is referred to as an accuracy measure which uses simulation to derive the impact of forecasting accuracy in terms of the inventory levels and the service levels achieved by the inventory system (Sani and Kingsman, 1997, Eaves and Kingsman, 2004). The above three groups of error measures are examined in this research.

2.6.1 Absolute measures of accuracy

Mean squared error (MSE) [equation (2-40)] is a commonly used absolute measure of accuracy for intermittent demand forecasts (Johnston and Boylan, 1996). MSE and root mean squared error (RMSE) [equation (2-41)] place a heavier weight on large errors (Gross and Sohl, 1990, Eaves, 2002). To avoid this effect, some research (Makridakis and Winkler, 1983, Dangerfield and Morris, 1992, Willemain et al., 1994, Ghobbar and Friend, 2003) has used mean absolute percentage error (MAPE) [equation (2-42)] as a unit-free measure. MAPE is useful when comparing different forecasts (Regattieri et al., 2005).

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \tag{2-40}$$

$$RMSE = \left(\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2\right)^{\frac{1}{2}} \tag{2-41}$$

$$MAPE = \left[\frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|\right] \times 100 \tag{2-42}$$

$$MPE = \left[\frac{1}{n}\sum_{t=1}^{n}\frac{y_t - \hat{y}_t}{y_t}\right] \times 100 \tag{2-43}$$

where:

y_t = the observed demand at time t

$\hat{y}_t$ = the estimated demand of y_t

$t = 1, 2, 3, \ldots, n$

n = the total number of time periods

For intermittent time series (i.e. many observations with zero or near-zero demand), MAPE [also mean percentage error (MPE)] are difficult to define because the percentage error in equations (2-42) and (2-43) cannot be calculated with a zero denominator (Gross and Sohl, 1990, Hyndman et al., 1998, Eaves and Kingsman, 2004, Regattieri et al., 2005, Syntetos and Boylan, 2005). *Ad hoc* procedures (e.g. excluding zero-demand time periods or adding a small amount to zero demand) could not contribute to significantly improving confidence (Syntetos and Boylan, 2005). For these reasons, mean absolute deviation divided by average (MAD/A) [equation (2-45)] (Regattieri et al., 2005, Boylan et al., 2008) or geometric root mean square error (GRMSE) [equation (2-47)] (Boylan et al., 2008) was used. MAD/A ignores its sign, taking its absolute values of errors, and summarises data across the time series by its mean value (Boylan et al., 2008). MAD is less sensitive to outliers than MSE (Eaves, 2002). Outliers are defined as "aberrant scores that lie outside the usual range of scores which are expected for a particular variable" (Miles and Shevlin, 2001, p. 63). GRMSE rests on squared errors and takes the appropriate geometric mean as the summary measure; thus the effect of high errors for outliers can be cancelled out (Boylan et al., 2008). Root mean square error (RMSE) has also been considered useful when measuring forecasting errors due to its simplicity and ability to weigh heavily the magnitude of errors (Gross and Sohl, 1990). As mentioned, MSE and RMSE place a heavier weight on large errors (Gross and Sohl, 1990, Eaves, 2002). Hence, MSE and RMSE are useful when large errors cause greater inventory costs in proportion to small errors (Kling and Bessler, 1985). As with RMSE, root mean squared error/average (RMSE/A) can be useful when large errors cause greater costs in proportion to small errors. MAD and MAD/A are also useful in order to avoid heavier weight on large errors. As such, MAD and MAD/A, together with RMSE and RMSE/A, can be utilised as error measures in order to capture the two alternative effects of larger errors upon weight (or costs).

$$MAD = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \hat{y}_t\right| \tag{2-44}$$

$$MAD / A = \frac{MAD}{\bar{y}} \tag{2-45}$$

$$RMSE / A = \frac{RMSE}{\bar{y}} \tag{2-46}$$

$$GRMSE = \left(\prod_{t=1}^{n}(y_t - \hat{y}_t)^2\right)^{\frac{1}{2n}} \tag{2-47}$$

where:

y_t = the observed demand at time t

$\hat{y}_t$ = the estimated demand of y_t

$\bar{y}$ = the average value of the demand time series y_t

$t = 1, 2, 3, \ldots, n$

n = the total number of time periods

On the other hand, Narasimhan et al. (1998) argued that unbiased forecasts should lead cumulative or a running sum of forecast errors (RSFE) to near zero. RSFE can be expressed as in equation (2-48). When the forecasts are continually high, negative

RSFE would be observed, and vice versa. A relative size of RSFE, S, can be expressed as in equation (2-49) (Narasimhan et al., 1998).

$$RSFE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t) \approx 0 \tag{2-48}$$

$$S = \frac{RSFE}{MAD} \tag{2-49}$$

S should be used with caution, because the purpose of S is to monitor the balance of negative and positive errors (Narasimhan et al., 1998). If negative errors and positive errors are cancelled each other out, a collection of very high negative errors and very high positive errors could result in a value close to zero. S close to zero does not mean an accurate forecast, but a balanced forecast.

2.6.2 Relative measures of accuracy

Assuming squared error, $e^2_{M,t}$, generated by a forecasting method, M, at time point t as in equation (2-50), method (M) specific error, $\varepsilon^2_{M,t}$ can be measured by a relative measure to another method (Syntetos and Boylan, 2005).

$$(y_t - \hat{y}_{M,t})^2 = e^2_{M,t} = \varepsilon^2_{M,t} u_t \tag{2-50}$$

where:

y_t = the observed demand at time t

$\hat{y}_{M,t}$ = the estimated demand of y_t by forecasting method M

$\varepsilon^2_{M,t}$ = method (M) specific error at time t

u_t = error due to the particular time point t affecting all methods equally

Syntetos and Boylan (2005) argued that the error can be contaminated by occasional outliers; however, relative geometric root-mean-square error (RGRMSE) is independent of the u_t and is a well-behaved relative accuracy measure for intermittent demand. RGRMSE for methods A and B in a time series is expressed as in equation (2-51).

$$RGRMSE = \frac{\left(\prod_{t=1}^{n}(y_t - \hat{y}_{A,t})^2\right)^{\frac{1}{2n}}}{\left(\prod_{t=1}^{n}(y_t - \hat{y}_{B,t})^2\right)^{\frac{1}{2n}}} \qquad (2\text{-}51)$$

where:

subscript A or B = the forecasting method A or B

$\hat{y}_{A,t}$ = the estimated demand of y_t by forecasting method A

Relative measures have been used to compare two forecasting method that use top-down and direct forecasting strategies. Gross and Sohl (1990) investigated a relative measure to compare top-down and direct forecasting methods at item level. They used composite RMSE differential (CD) to summarise relative performance of forecasting strategies. RMSE for a top-down forecasting method ($RMSE_{TD}$) which can be calculated as in equations (2-52) and (2-53). CD can be calculated as in equation (2-55). A negative score of CD suggests that a top-down forecasting method provides smaller CD than a direct forecasting method; a positive score of CD indicates that a direct forecasting method provides smaller CD than a top-down forecasting method.

$$f_{i,t} = PR_{i,t}F_t \qquad (2\text{-}52)$$

$$RMSE_{TD(i)} = \left(\frac{1}{n} \sum_{t=1}^{n} (y_{i,t} - f_{i,t})^2 \right)^{\frac{1}{2}} \tag{2-53}$$

where:

$f_{i,t}$ = the forecast of item i at time point t, using $PR_{i,t}$ as the proration method

$PR_{i,t}$ = the proration method used to allocate item i's share of the total forecast at time point t

F_t = the total forecast for all products in the line at time point t

$RMSE_{TD\,(i)}$ = RMSE of a top-down forecasting method for item i

$y_{i,t}$ = the observed demand of item i at time point t

$t = 1, 2, 3, \ldots, n$

n = the total number of time periods

$$D_i = RMSE_{TD(i)} - RMSE_{DF(i)} \tag{2-54}$$

$$CD = \sum_{i=1}^{N} D_i \overline{PR_i} \tag{2-55}$$

where:

D_i = RMSE differential

$RMSE_{DF(i)}$ = RMSE of a direct forecasting method for item i

CD = composite root mean square error differential

$\overline{PR_i}$ = average proportion of total group demand for item i

$i = 1, 2, 3, .., N$

N = the total number of items within the group

Widiarta et al. (2008b) used the proportion of RMSE, Δ, as a relative error measure [equation (2-56)]. A value of $\Delta < 1$ indicates that a top-down forecasting method outperforms a direct forecasting method, and a value of $\Delta > 1$ indicates that a direct forecasting method outperforms a top-down forecasting method

$$\Delta = RMSE_{TD}/RMSE_{DF} \tag{2-56}$$

Dangerfield and Morris (1992) proposed a relative measure. In the first step, they produced a relative direct summary measure as with Widiarta et al. (2008b); that is, $MAPE_{TD}$ / $MAPE_{DF}$. In the second step, the natural log of the ratios of errors for the two forecasts was calculated as in equation (2-57). A positive log relative error indicates that a direct forecasting method outperforms top-down forecasting method, and vice versa. Dangerfield and Morris (1992) suggested that this measure is an unbiased relative measure. For example, assuming that the $MAPE_{TD}$ and $MAPE_{DF}$ of a time series "A" are 2 and 1, and the $MAPE_{TD}$ and $MAPE_{DF}$ of a times series "B" are 1 and 2, the simple ratio, $MAPE_{TD}$ / $MAPE_{DF}$ of the time series, "A" and "B", are 2.0 and 0.5 respectively. Thus, the total average ratio is 1.25 = (2 + 0.5) / 2, although the overall performance of the two methods are identical. However, as the average natural log of the ratios of errors for the same time series, "A" and "B", which are 0.69 and -0.69 respectively, is 0 = (0.69 – 0.69) / 2, this can furnish an unbiased summary statistic.

$$\textit{Log relative error} = \ln(MAPE_{TD} / MAPE_{DF}) \tag{2-57}$$

2.6.3 Derivative measures of accuracy

There are studies which have considered the derivative impact of forecasting methods on stock-holding (Johnston and Boylan, 1996, Boylan et al., 2008). Two kinds of comparisons were carried out. Firstly, errors were summed at every point in time;

secondly, the comparisons were made only immediately after an issue has occurred. The authors (Johnston and Boylan, 1996, Boylan et al., 2008) argued that the first comparisons would be associated with a re-order interval stock replenishment system; whereas the second would correspond to a re-order level stock replenishment system. The re-order interval stock replenishment system which is also known as a periodic review system is referred to as an inventory control system that defines time points for examining inventory levels and then makes decisions accordingly. The re-order level stock replenishment system which is also known as a continuous review system is referred to as an inventory control system that makes inventory-related decisions when inventory reaches a particular level (Slack et al., 2004).

Sani and Kingsman (1997) investigated a forecasting accuracy measure scheme which assesses the regret of using a particular method compared to the best method for a specific time series. They compared forecasting methods and inventory methods in terms of either average inventory cost regret (RC) [equation (2-59)] or average service level regret (RS) [equation (2-61)]. The ratio of RC and RS is expressed as a percentage.

$$C_i = \frac{1}{5}\sum_{k=1}^{5} C_{i,k} \tag{2-58}$$

$$RC_i = \frac{1}{Mn_i}(C_i - Mn_i) \tag{2-59}$$

where:

C_i = the average inventory cost for item i over the five forecasting methods

$C_{i,k}$ = the inventory cost for item i using forecasting method k

$Mn_i = min(C_i)$, the minimum cost for item i across the inventory methods

$$S_i = \frac{1}{5}\sum_{k=1}^{5} S_{i,k} \tag{2-60}$$

$$RS_i = \frac{1}{MX_i}(Mx_i - S_i) \tag{2-61}$$

where:

S_i = the average service level for item i over the five forecasting methods

$S_{i,k}$ = the service level for item i using forecasting method k

Mx_i = $max(S_i)$, the maximum service level

Eaves and Kingsman (2004) attempted to measure forecasting accuracy in terms of implied stock-holding. Average stock-holdings using each forecasting method were computed and compared through a backward-looking simulation. The backward-looking simulation is referred to as an inventory control simulation using historical real data. A periodic review inventory system was simulated. A forecasting method which leads to the lowest stock-holding was regarded as the best forecasting method. This measure was argued to provide more intuitive comparison results compared to other measures because this can present forecasting performance in terms of inventory carrying costs (Eaves and Kingsman, 2004).

In this section, various accuracy measures were reviewed. Derivative measures were argued to present practical and intuitive comparisons. In order to implement a derivative measure, a review of theories related to inventory system is necessary. In the next section, inventory theories for implementing derivatives measures are reviewed.

2.7 Derivative Measures of Inventory Model Performance

The first issue involved in an inventory system might be establishing how critical the item under consideration is to the organisation (Silver et al., 1998). This section reviews theories related to inventory classification which establishes the inventory priorities. This is followed by a review of inventory system control theories. Then, inventory model measurements are examined.

2.7.1 Inventory classification

For the purposes of inventory control, ABC classification can be utilised. ABC classification can be described as: a) A: expensive items needing special care; b) B: ordinary items needing standard care; c) C: cheap items needing little care (Silver et al., 1998, Waters, 1991). More concentration is placed on higher value items than on lower value items. This analysis is sometimes called a Pareto analysis, or the 'rule of 80/20' (suggesting that 80% of inventory items need 20% of the attention, while the remaining 20% of items need 80% of the attention) (Waters, 1991). An ABC classification does not have to be done on the basis of the classification of value alone. Managers may consider other reasons such as criticality of the operation to the firm (Silver et al., 1998).

Table 2-8 Weapon System Indicator Code (Deshpande et al., 2003)

Part essentiality code / Mission criticality code	Very High	High	Medium	Low
High	A	B	C	D
Medium	E	F	G	H
Low	I	J	K	L

The US Defence Logistics Agency (DLA) decides inventory investments to balance two conflicting goals: minimising holding and investment costs while maximising mission readiness (Deshpande et al., 2003). The importance of parts in supporting readiness is measured by two dimensions: essentiality to the operation of the weapon system in which it is housed, and criticality of that system to the user's overall mission. A part is assigned a Weapon System Indicator Code (WSIC) as shown in Table 2-8. As such, the DLA differentiates the service level for different WSICs (Deshpande et al., 2003).

2.7.2 Inventory systems

It is necessary to define inventory systems and terminologies related to inventory systems before discussing inventory system theories. Silver et al. (1998, pp. 233 - 234) defined terminologies related to inventory control systems as:

a) Stock on hand is a stock which is physically on the shelf;
b) Stock on order is a stock which has been requisitioned but not yet received;
c) Stock-out is an occasion when the stock on hand drops to the zero level;
d) Net stock can be calculated as "net stock = stock on hand – backorders"; and
e) Inventory position can be calculated as "inventory position = (stock on hand) + (stock on order) – (backorders)".

Slack et al. (2004) divided inventory systems into single-stage, two-stage, multi-stage, or multi-echelon inventory systems by the positions of inventories:

a) A single-stage inventory system has only one stock position to manage;
b) A two-stage inventory system has a central depot and various local distribution points;
c) A multi-stage inventory system has various stages of stocks as a form of work-in-process in which materials from suppliers flow through the various stages of a production process; and

d) Multi-echelon inventory systems refers to interconnected sets of inventory systems in which materials flowing through the inventory systems are stored at different points before reaching customers.

Two approaches to the decision on when to place an order (i.e. the continuous and periodic review systems) were introduced in Subsection 2.6.3. These can be expounded as:

a) The continuous review system (fixed order quantity system) processes reviewing inventory status continuously and then places an order as a constant order quantity (Q) when the inventory position reaches its re-order point (s) (Waters, 1991, Bowersox and Closs, 1996, Waller, 2003, Slack et al., 2004). The re-order point is determined by estimating the expected usage of inventory during lead time, plus a safety stock (Waller, 2003). In order to process this system effectively, accurate calculation and computer assistance is required (Bowersox and Closs, 1996).
b) The periodic review system (fixed time period system) places orders of variable quantities at regular intervals (R) (Waters, 1991, Bowersox and Closs, 1996, Waller, 2003, Slack et al., 2004). An order quantity is determined to cover demand between the replenishment order being placed and the following replenishment order arriving.

Silver et al. (1998) discussed the advantages and disadvantages of the continuous and periodic review systems. The periodic review system can allow a reasonable prediction of the level of the workload and also allow the same review interval on every member in a supply chain so that coordination of replenishment can be achieved. However, this system is likely to involve lager stock holdings than the continuous review system. Under the continuous review system, the workload is less predictable because a replenishment decision can be made at any moment and more costs for the reviewing process are required. However, the continuous review system can provide the same level of customer service continuously and requires less safety stock than does the periodic review system. When demand is unusually high between ordering times under the periodic review system, stock-out is more likely to occur (Waller, 2003).

To accommodate specific situations, the variations of the periodic and continuous review systems have been developed (Bowersox and Closs, 1996, Silver et al., 1998). An order-point, order-quantity (s, Q) system, also known as two-bin system, is a special form of the continuous review system (Silver et al., 1998, Waller, 2003, Slack et al., 2004). A constant order quantity (Q) is ordered when the inventory position reaches the reorder point (s). A simple two bin system stores the re-order level inventory plus the safety inventory in the second bin whilst using spare parts from the first bin. When the spare parts in the first bin are depleted, a new re-order is placed and the spare parts in the second bin are used. When the replenishment arrives, the second bin is refilled and the remainder moves into the first bin. An inventory system which has a third bin for storing the safety inventory separately refers to a three-bin system. The (s, Q) system is quite simple, so that errors are less likely to occur and production requirements are predictable for suppliers. However, there is a disadvantage that its inflexible order quantity is unlikely to cope with unusual large demand (Silver et al., 1998).

In contrast to the (s, Q) system, a continuous review system uses a variable replenishment quantity, which raises inventory position to the order-up-to-level (S). An order-point, order-up-to-level (s, S) system, also known as min-max system or optional system, is another special form of the continuous review system (Bowersox and Closs, 1996, Silver et al., 1998). The best (s, S) system could have inventory and shortage smaller than those of the best (s, Q) system. However, there are no simple procedures nor algorithms to obtain the best values of the control parameters, s and S, in any practical situation (Sani and Kingsman, 1997, Silver et al., 1998). In the (s, S) system more errors are likely to be made by suppliers for the variable, order quantity, than the (s, Q) system (Silver et al., 1998).

A periodic-review, order-up-to-level (R, S) system, also known as a replenishment cycle system, is a special form of the periodic review system (Bowersox and Closs, 1996, Silver et al., 1998). At every review interval (R), enough order quantity is ordered to raise inventory position to the order-up-to-level (S). This system has the advantage of coordination of replenishment. However, this system has the disadvantage of larger stock holdings than the continuous review system (Silver et al., 1998).

A combination of the (s, S) system and the (R, S) system refers to the (R, s, S) system (Silver et al., 1998). At every review interval (R), the inventory position is checked. If the inventory position reaches the reorder point (s), enough order quantity is ordered to raise inventory position to the order-up-to-level (S). Otherwise, no order is placed until the next review. The best (R, s, S) system could produce lower inventory carrying costs and lower inventory stock-out costs than the above systems, however, it is difficult to obtain the best control parameters, R, s, and S (Silver et al., 1998). Owing to the difficulty of obtaining the best control parameters, Silver et al. (1998) suggested that the (s, S) and (R, s, S) systems are used for A items, and the (s, Q) and (R, S) systems are used for B items.

There is research that compared the performance of the (s, S) system and the (R, S) system for spare parts demand. Sani and Kingsman (1997) examined the performance of the various forms of the (s, S) system and the (R, S) system with 30 daily spare parts demand data for vehicles and agricultural machinery over five years. They classified the spare parts into three groups by their annual demand and compared the inventory systems in terms of either average inventory cost regret (RC) or average service level regret (RS) as shown in Table 2-9. In terms of average inventory cost regret (RC), compared with the (s, S) systems, the (R, S) system performed reasonably well for the very low demand items, but badly for the medium and high demand items. In terms of average service level regret (RS), the (R, S) system provided poor performance, compared with the (s, S) systems.

Table 2-9 The performance of inventory systems for spare parts

(Sani and Kingsman, 1997)

Group	Very low demand		Low to medium demand		High demand	
Annual demand	Less than 20 units		Between 20 and 40 units		Over 40 units	
Measure	RC	RS	RC	RS	RC	RS
The best (s, S)	8%	3%	2%	2%	2%	1%
The worst (s, S)	72%	10%	61%	17%	63%	8%
The (R, S)	19%	12%	52%	15%	60%	18%

RC = average inventory cost regret [equation (2-59)]; RS = average service level regret [equation (2-61)].

2.7.3 Measures

Theories for inventory systems were reviewed above. These theories could provide the basis for the assessment of the performance of an inventory system to derive the impact of a forecasting method on the inventory system. Clarification of an appropriate measurement to assess the performance of an inventory model is another important issue related to the derivative measure of accuracy.

The performance of an inventory model can be measured by the total inventory costs and the inventory fill rate (Petrovica et al., 1998). This subsection reviews measurements for inventory costs. This is followed by measurements for customer service including the inventory fill rate.

Costs relevant to the inventory management can be categorised as five factors: unit variable cost, inventory carrying costs, inventory ordering costs, inventory stock-out costs, and system control costs (Silver et al., 1998):

a) Unit variable cost of an item can be expressed in a monetary value per unit. This can be merely the price paid to the supplier.
b) Inventory carrying costs comprise the opportunity costs of the money invested, warehousing costs, and special holding costs such as insurance and obsolescence.
c) Inventory ordering costs are those related costs to procurement.
d) Inventory stock-out costs are those relevant costs with insufficient inventory to satisfy user demand. This includes backordering costs and the costs of lost demand.
e) System control costs are those related costs with the operation of the specific decision system such as data acquisition, data storage and computation.

In practice for the purpose of measuring the performance of an inventory model, many researchers calculate the total inventory costs as the sum of inventory carrying costs and inventory stock-out costs (Sterman, 1989, Sani and Kingsman, 1997, Petrovica et al., 1998, An et al., 2002). Sterman (1989) quantified weekly inventory carrying costs and weekly inventory stock-out costs per item as \$0.5 and \$1.0 respectively for the "beer distribution game". Sani and Kingsman (1997) quantified annual inventory carrying costs and annual inventory stock-out costs per item as 20% and 33% of the unit variable cost respectively for spare parts for vehicles and agricultural machinery. Petrovic et al. (1998) conducted a supply chain simulation with generated demand sets, and specified unit carrying costs as 1.5, 1, 0.5, 0.3, and 0.1, and unit stock-out costs as 6.5, 4.3, 2.15, 1.29, and 0.43 for five products. For a spare parts inventory system simulation for tanks in the South Korean Army, An et al. (2002) quantified inventory carrying costs and inventory stock-out costs per item as \$1.0 and \$2.0 respectively.

Slack et al. (2004) noted five strategic roles of inventory. They are supporting quality objectives, speed objectives, dependability objectives, flexibility objectives, and cost objectives. In militaries, supporting dependability objectives (i.e. sustaining operational availability) and supporting speed objectives (i.e. fast responding to the demand) can be more important than supporting cost objectives because of the seriousness of a result which could be caused by the unavailability of a weapon system. This could lead militaries to large stock holdings. Large stock holdings amplify the disadvantages of stock holding such as obsoleteness, damage, deterioration and loss (Slack et al., 2004). In practice, as stated in Chapter 1, in many militaries the ability to stock spare parts is constrained by limited budgets. Redundant large stock holdings as well as stock-out are to be avoided. Therefore, both inventory carrying costs and inventory stock-out costs might need to be considered to measure the performance of a military inventory model.

Waller (2003) claimed that the inventory carrying costs are up to one-third of the value of the inventory. He pointed out that inventory stock-out costs are difficult to quantify directly. Although quantifying the inventory stock-out costs are difficult in business, they can be quantified by backorder costs, lost sale costs, or lost customer costs (Waller, 2003). However, it is more difficult in militaries, because the stock-out costs could lead to a military defeat that could cause casualties and deaths (MacDonald, 1997).

In order to avoid calculating the difficult inventory stock-out costs, a safety margin (i.e. safety stock), which provides inventory to user units as a fill rate of one hundred percent, was introduced (Wemmerlöv, 1989, Eaves, 2002, Eaves and Kingsman, 2004). They measured the performance of an inventory system only by inventory-holdings or inventory carrying costs. Eaves (2002) and Eaves and Kingsman (2004) used the implied stock-holdings to measure the performance of the inventory simulation model for spare parts in the UK Air force. With an (R, S) system model the safety stock levels were iteratively added to the order-up-to-level (S), until no stock-outs occurred. To achieve one hundred percent fill rate, an average of ten iterations was carried out for each item. As such, the implied stock-holdings (i.e. the measure of the inventory system) were calibrated as the average of the opening stock plus deliveries and the closing stock (Eaves, 2002, Eaves and Kingsman, 2004). The implied stock-holdings were also converted into the monetary values of the additional investment in stock-holdings so that the amount of additional stock-holdings was quantified in terms of costs.

The performance of an inventory model can also be measured by customer service. Silver (1970) addressed three measures of customer service such as the item availability, the fraction of demand satisfied without backorder, and the average amount on backorder at a random point in time. The item availability is referred to as the fraction of time that the stock on hand is greater than zero (Silver, 1970). Silver et al. (1998) developed these three customer service measures into four customer service measures such as the cycle service level, the ready rate, the average time between stock-out (TBS) occasions, and the fill rate. The cycle service level is referred to as the fraction of replenishment cycle in which a stock-out does not occur. The ready rate is referred to as the fraction of time during which the net stock is positive. TBS can be calculated as the reciprocal of the average number of stock-out occasions per year. The fill rate is referred to as the fraction of customer demand which is met routinely without backorders or lost sales. The inventory fill rate can be computed as the ratio of demand immediately filled from the stock on hand. A simple equation of the fill rate can be expressed as (Heuts et al., 1999):

$$Fill\ rate = 1 - \frac{mean\ shortage}{mean\ demand} \tag{2-62}$$

where: $shortage = demand\ quantity - (stock\ on\ hand + delivery\ quantity)$.

2.8 Summary and Conclusion

Traditionally, various direct forecasting methods such as exponential smoothing, weighted moving average, Croston's method, Syntetos-Boylan approximation, and the Box-Jenkins models were considered as appropriate forecasting methods for non-normal demand. However, hierarchical forecasting might be applicable to military spare parts demand which is likely to be non-normal. There are three sub-strategies of hierarchical forecasting, namely bottom-up forecasting, top-down forecasting, and combinatorial forecasting. Research found that top-down forecasting presents inferior performance to that of direct forecasting or conditional performance to some demand features. Combinatorial forecasting was suggested to be a superior forecasting strategy to other forecasting strategies (Kahn, 1998, Dekker et al., 2004, Hyndman et al., 2007). Lack of agreement in the performance of hierarchical forecasting, which might originate from the different features of data, was identified. Therefore, the influence of demand features upon the performance of forecasting methods was reviewed.

Classification scheme for non-normal demand and its influence upon the performance of direct forecasting methods are rather well-developed. Various demand features, such as average inter-demand interval, coefficient of variation in demand size, number of periods with zero demand, lead time variability, trend, number of peaks, and skewness, have been examined to classify non-normal demand. Among the studies, Johnston and Boylan (1996) and Boylan et al. (2008) developed practical classification schemes which present the boundary values of the non-normal demand features for direct forecasting methods.

Literature about the influence of demand features upon the performance of hierarchical forecasting was also reviewed. Demand features, which were considered to influence the performance of top-down forecasting, are correlations, proportion, variability, the degree of substitutability and the variability of an item's proportion, forecasting horizon, lag-1 autocorrelations, grouping criteria, and seasonality.

Measurement of forecast accuracy is important, because different accuracy measures can lead to different conclusions. Accuracy measures are of three groups, that is, absolute measures, relative measures, and derivative measures. Absolute measures, such as MAD, MAD/A, RMSE, RMSE/A, and a running sum of forecast errors divided by MAD (RSFE/MAD) are useful, when evaluating a forecasting method in isolation for non-normal demand. As a relative measure, the log relative error is useful when comparing alternative two forecasting methods across a set of time series. A derivative measure uses simulation to derive the impact of forecasting accuracy in terms of the inventory levels and the service levels achieved by the inventory system. For the purpose of implementing the derivative measure theories related to the measurement of the performance of an inventory model were also reviewed.

This chapter identified two research gaps:

a) Little attention has been paid to the use of hierarchical forecasting for the intermittent demand at item level which is a feature of non-normal demand associated with spare parts demand. In spite of its high applicability, no research has examined the applicability of hierarchical forecasting to intermittent demand at item level.
b) There has been little discussion about the influence of non-normal demand features upon the performance of hierarchical forecasting. No research has investigated the influence of correlations between non-normal time series upon the relative performance of hierarchical and direct forecasting methods. There has been no controlled research which has examined the influence of demand features upon the

performance of combinatorial forecasting. No research has found a combined influence of non-normal demand features upon the relative performance of hierarchical forecasting.

The first research gap leads to research questions a), b), and c). It can be assumed on the basis of the existing literature that spare parts demand in the South Korean Navy is non-normal (Markland, 1970, Businger and Read, 1999, Eaves and Kingsman, 2004). However, the nature of the spare parts demand in the Navy is required to be identified empirically. This requirement leads to research question a) "what is the nature of the spare parts demand in the South Korean Navy?" When the nature of the spare parts demand is identified, it is also expected that hierarchical forecasting can be superior for the spare parts demand to direct forecasting (DeLurgio, 1998, Hyndman et al., 2007, Widiarta et al., 2008b). This expectation leads to research question b) "what forecasting method is appropriate for the spare parts demand in the South Korean Navy?" Authors argued that combinatorial forecasting is a superior forecasting strategy to other strategies (Kahn, 1998, Dekker et al., 2004, Hyndman et al., 2007). Once the nature of the spare parts demand is identified, the superiority of combinatorial forecasting for the spare parts demand is investigated. This enquiry leads to research question c) "under what conditions are top-down forecasting or combinatorial forecasting superior or inferior to direct forecasting?"

As results of answering research questions a), b), and c) with respect to the first research gap, the following contributions will be achieved: the nature of the spare parts demand in the South Korean Navy will be identified and the performance of competing forecasting strategies (hierarchical forecasting and direct forecasting) for the spare parts demand will be also identified.

The second research gap leads to research questions c) and d). The second research gap requires research on the influence of correlations between the spare parts demands upon the performance of combinatorial forecasting. Research on a combined influence of multiple demand features including correlations and intermittency upon combinatorial forecasting is also required. These requirements lead to research question c) "under what conditions are top-down forecasting or combinatorial forecasting superior or inferior to direct forecasting?" Furthermore, research on a multivariate classification model that predict the relative performance of alternative forecasting methods (hierarchical and direct forecasting methods) for spare parts demand by the multivariate demand features including correlations and intermittency is required. This requirement leads to research question d), "how can the spare parts demand be classified in order to predict a superior forecasting method?"

As results of answering research questions c) and d) with respect to the second research gap, the following contributions will be achieved: the influence of demand features upon combinatorial forecasting will be identified; a new classification model for the spare parts demand which predicts the relative performance of the alternative forecasting methods by the multivariate demand features will be developed; and the research findings will be validated with diagnostics, cross-validation and a variety of accuracy measures including derivative measures using simulation with empirical data.

In this chapter, literature related to this research was reviewed and the research gaps in the literature were identified. It was demonstrated that the research questions were derived from the research gaps. Then, the contributions which will be achieved as results of answering the research questions were identified. After reviewing literature, the next step antecedent to collecting data is designing the research (Saunders et al., 2007). The next chapter provides a methodological framework which will be applied in this research.

Chapter 3. Methodology

This chapter delineates the research methodology. It starts by identifying the purpose of this research. In Section 3.2, the methodological categories in operational management research are reviewed. In Section 3.3, the suitability of the case study to the research questions is discussed. In Section 3.4, the relationship of theory to the case study is examined. In Section 3.5, the case study research design is outlined. In Section 3.6, the criteria for evaluating the research are discussed. In Section 3.7, the research procedure of this study is presented. Finally, this chapter is summarised in Section 3.8.

3.1 Purpose of Research

The purposes of research can be categorised as exploratory, descriptive and explanatory (Saunders et al., 2007). An exploratory study can be described as finding out what is happening and asking questions and assessing phenomena in a new light. A descriptive study can be described as portraying an accurate profile of persons, events or situations. An explanatory study can be described as establishing causal relationships between variables.

Recalling the aim of this research, which is "to establish an appropriate forecasting strategy for predicting the demand for spare parts in the South Korean Navy", this can be achieved by answering the research questions. This requires ascertaining the nature of spare parts demand and the forecasting system in the South Korean Navy and identifying an appropriate forecasting method for the spare parts demand in the Navy with a new forecasting strategy (i.e. hierarchical forecasting strategy) and a new measurement (i.e. derivative measure of accuracy). This also requires establishing the relationships between forecasting methods and the performance of the forecasting methods, and between demand features and the performance of the forecasting methods. Therefore, this research might be in line with an exploratory study as well as an explanatory study.

3.2 Empirical Research in Operations Management

Operations management is defined as "the effective planning, organising, and control of all resources and activities necessary to provide the market with tangible goods and services. It applies to manufacturing, service industries and *not-for-profit organisations*" (Waller, 2003, p. 875). This research focuses on model-based quantitative research in operations management disciplines applied to a *not-for-profit organisation* (i.e. the South Korean Navy). Model-based quantitative research refers to research where models of causal relationships between control variables and performance variables are developed, analysed or tested (Bertrand and Fransoo, 2002). In causal relationships, a change of value α in one variable will lead to a change of $f(\alpha)$ in another variable, so that a model can be utilised to predict the future state of the modelled processes (Bertrand and Fransoo, 2002). In this research, the future state refers to the change of performance variables such as inventory level, inventory costs, and inventory fill rate. The change of forecasting methods for spare parts demand can lead to a change in the performance variables.

Bertrand and Fransoo (2002) classified quantitative operations management research as either axiomatic research or empirical research: a) axiomatic quantitative research indicates the process of achieving resolutions by the defined model; b) empirical quantitative research indicates the process of achieving resolutions by empirical findings. Bertrand and Fransoo (2002) divided axiomatic and empirical quantitative research into two sub-categories by their objectives, namely normative research and descriptive research: normative research has the objective of establishing policies, strategies and actions; descriptive research has the objective of analysing models or describing the causal relationships in reality.

In terms of the classification scheme from Bertrand and Fransoo (2002), this research might be in line with empirical normative quantitative research. This is because the resolutions were achieved by empirical findings; and the aim of this research is to establish an appropriate forecasting strategy for predicting the demand for spare parts in the South Korean Navy. Empirical normative quantitative research was claimed to be difficult to verify because controlling all relevant variables is impossible and this is a requirement for evaluating performance changes in empirical normative quantitative research (Bertrand and Fransoo, 2002). The verification of the performance of a forecasting method is difficult because different forecasting accuracy measures can lead to different conclusions as stated in Section 2.6 (Syntetos and Boylan, 2005). This research compares the performance of different forecasting methods in the context of spare parts demand in the South Korean Navy. In order to assess the practical impact that a forecasting method has on the inventory system, controlling all relevant variables such as the forecasting review cycle and the procurement lead times and purchasing prices of spare parts is required. Thus, the verification is an important issue for this research.

Wacker (1998) classified operations management research as either analytical (formal) research or empirical research for the purpose of theory building as shown in Table 3-1. In empirical statistical research, theoretical relationships are verified statistically in large external samples from reality. However, an empirical case study looks into small samples to test and develop complex relationships between variables to suggest a new theory (Wacker, 1998).

In terms of the classification scheme from Wacker (1998), this research might be in line with an empirical quantitative case study because this research investigates only one case (i.e. the case of spare parts in the South Korean Navy) to test and develop a forecasting strategy.

Table 3-1 Research category in operations management (Wacker, 1998)

		Types of research included
Analytical	Conceptual	Futures research scenarios, introspective reflection, hermeneutics, conceptual modelling
	Mathematical	Reason/logical theorem providing normative analytical modelling, descriptive analytical modelling, proto-typing, physical modelling, laboratory experiments, mathematical simulation
	Statistical	Mathematical statistical modelling
Empirical	Experimental design	Empirical experimental design, descriptive analytical modelling
	Statistical sampling	Action research, structured and unstructured research, surveying, historical analysis, expert panels
	Case studies	Field studies, case studies

3.3 Case Study Strategy

A case study is defined as "an empirical inquiry that investigates a contemporary phenomenon within its real-life context" (Yin, 2003, p. 13). A case study is a research strategy which concentrates on perceiving the dynamics present within single settings (Eisenhardt, 1989). Saunders et al. (2007, p. 139) noted that a case study strategy is most often used in explanatory and exploratory research. This research has been stated to be exploratory study as well as explanatory study. A case study is suitable for achieving these research purposes.

A case study is particularly good for examining "why" as well as "how" and "what" questions among question series, "who", "what", "where", "how" and "why", which are enquiries about a contemporary set of events over which the investigator has little or no control (Yin, 2003, Saunders et al., 2007). A case study is especially suitable for "how" questions because these questions deal with operational links needed to be traced over time, rather than mere frequencies or incidence (Yin, 2003). Table 3-2 restates the research questions of this research. "How" and "what" questions are noticed as shown. A case study is appropriate to answer these questions.

Table 3-2 The research questions

a) What is the nature of the spare parts demand in the South Korean Navy? b) What forecasting method is appropriate for the spare parts demand in the South Korean Navy? c) Under what conditions are top-down forecasting or combinatorial forecasting superior or inferior to direct forecasting? d) How can the spare parts demand be classified in order to predict a superior forecasting method?

3.4 Theory in Case Study

A theory is defined as "a statement of relationships between units observed or approximated in the empirical world" (Bacharach, 1989, p. 498). A case study can be used for theory building, theory testing, or theory refinement (Voss et al., 2002). This section examines the association between theory and the findings of research and the issue of theory induction from data.

3.4.1 Theory as knowledge base

Research is based on theory as a knowledge base. There are three possibilities in which: a) existing theory provides no framework for findings; b) existing theory conflicts with findings; and c) existing theory is in accordance with findings (Eisenhardt, 1989, McCutcheon and Meredith, 1993, Yin, 2003). When the existing theories are indigent and the available literature provides no conceptual framework or hypotheses of note, such a knowledge base may not be a good theoretical foundation, and any new empirical study is likely to assume the characteristic of an "exploratory" study (Yin, 2003).

On the other hand, the existing theories can be either in accord with the findings or in disagreement with the findings; in both cases, theory is important (Eisenhardt, 1989). Eisenhardt (1989) noted two reasons why enfolding literature which conflicts with the emergent theory is important: a) if conflicting findings are ignored, then confidence in the finding is reduced; b) conflicting literature generates an opportunity for refining the theory.

Theories for a case study can be very well-developed, especially where the object is to test or compare theories against empirical evidence, or the necessary foundation is available in the well-developed theories from other disciplines (McCutcheon and Meredith, 1993). Such cases are also important, because these theories bind together underlying similarities in phenomena typically not related to each other (Eisenhardt, 1989). This process of linking results might be essential in a theory-building case study research because the findings often rely upon a very restricted number of cases (Eisenhardt, 1989).

Theory is important in this research, in that the findings of this research are related to existing literature. However, as discussed in Subsection 2.2.3, theories related to the hierarchical forecasting for non-normal demand associated with spare parts demand are not well-developed. As such, two research gaps were identified. However, the necessary theoretical base is available from literature for top-down forecasting strategy as well as direct forecasting strategy in similar contexts.

3.4.2 Theory induction from data

A theory can be formed by either induction or deduction (Saunders et al., 2007). Wacker (1998) pointed out that the pivotal distinction between a case study and an analytical method is that empirical case study methods use induction (i.e. depend on data) and analytical methods use deduction. If a theory is based on data, then a large amount of data is required and case studies are a prime source of this kind of research (McCutcheon and Meredith, 1993). The data can be quantitative or qualitative and they can be collected from either single or multiple cases (Yin, 2003). This research adopted an empirical case study strategy employing an inductive method.

3.4.3 Generalisation

Eisenhardt (1989) argued that binding the emergent theory with existing literature strengthens the internal validity, generalisability (external validity), and theoretical level of theory building from case study research. Internal validity demonstrates a causal relationship, in which certain conditions lead to other conditions; and external validity tests whether a study's findings are generalisable beyond the immediate case study (Yin, 2003). Generalisability is a particular concern for a single case study as in this research (Saunders et al., 2007). In this case, analytic generalisation can be claimed for the case study research (Yin, 2003). There are two kinds of generalisation from case to theory: statistical generalisation and analytic generalisation (Yin, 2003).

In statistical generalisation, generalisability is established by an inference made about a population on the basis of empirical data collected about a sample (Yin, 2003). However, statistical generalisation should not be considered to be the method of generalising the results of the case study (Yin, 2003).

In analytic generalisation, generalisability is established by the process as: an existing theory is used as a framework with which to collate the empirical results of the case study; then, when more cases support the same theory, replication can be claimed (McCutcheon and Meredith, 1993, Yin, 2003). Analytic generalisation can be used in either single or multiple case study (Yin, 2003).

This research employed analytic generalisation in single case study design. A single case study (i.e. the case of spare parts in the South Korean Navy) was used for advocating or refining existing theories. Then, the theory established from the case study could extend to other situations such as other militaries and business logistics.

3.5 Research Design

Research design is the logical sequence that links the empirical data to a study's initial research questions; that is, the design discourages the situation in which the evidence is disconnected from the initial research questions (Yin, 2003). In this section, four components of research design are considered. Then, the rationales of the single case design for this research are presented.

3.5.1 Four components of research design

Yin (2003, p. 21) identified five components of a case study research design: a) a study's questions; b) its proposition, if any; c) its unit(s) of analysis; d) the logic linking the data to the propositions; and e) the criteria for interpreting the findings. Components a), b) and c) refer to what data are to be collected, whereas components d) and e) refer to what is to be done after the data have been collected (Yin, 2003). In this research, the four components a), c), d) and e) were considered because this research did not make a research proposition.

The first component is research questions. Although a case study is an inductive approach, a preliminary view of the general constructs or categories and their relationships, is required; then, initial research questions behind the proposed study should be followed (Voss et al., 2002). Even though the prior questions are tentative, it is crucial to establish a well-defined focus at the start, and to direct the collection of data (Voss et al., 2002). The initial research questions of this research directed this research to focus upon the research topic, and review the related literature. After reviewing the literature, the research gaps were identified. Then, the research gaps have led to the genuine research questions as shown in Table 3-2, and guided to collect the spare parts data in the South Korean Navy.

The second component, the unit of analysis is relevant to the fundamental problem of defining what the 'case' is (Yin, 2003). Precisely specifying research questions leads into the appropriate selection of the unit of analysis (Yin, 2003). As shown in Table 3-2, the research questions lead to the one unit of analysis; that is, the case of spare parts in the South Korean Navy. Once a general definition of the case has been established, other clarifications in the unit of analysis become important: for example, specific group of people, district boundary, or specific time boundary (Yin, 2003). Specifically three types of warships were clarified in this research. This is because these three types of warships have consumed a large volume of spare parts and many of these warships use the same pieces of equipment. The time boundary was defined from January 2002 to November 2007. This is because the Naval maintenance data system, which is the major data source, has been stabilised since 2002.

The third component, linking data to propositions, is a way of relating the data to the propositions (Yin, 2003). In lieu of the proposition, research questions were considered to be the objectives to link the data to. In order to relate the data to research question a), the spare parts demand and the current inventory control methods of the South Korean Navy are analysed in Chapter 4; to research question b), forecasts are generated with the data by the competing forecasting methods and compared with each other in Chapter 5 and 6; to research question c), the performance of the competing forecasting methods under different conditions such as accuracy measures, equipment groups and demand features are examined in Chapters 5 and 6; and to research question d), the relative performance of the alternative forecasting methods (hierarchical and direct forecasting methods) in the classification model is examined in Chapter 6.

The last component is the criteria for interpreting the findings. The results of tests which adopt competing forecasting methods are interpreted using appropriate accuracy measures (i.e. absolute and relative measures) in Chapters 5 and 6. This is verified by derivative measures using simulation.

3.5.2 Single case design

Voss et al. (2002) discussed case study research design in terms of the number of cases. They categorised it as either single cases or multiple cases, and illustrated the advantages and disadvantages of each as shown in Table 3-3. Single cases have the advantage of greater depth. In this research, the single case design (i.e. the case of spare parts in the South Korean Navy) was expected to provide more opportunity for in-depth observation.

Table 3-3 Choice of number of cases (Voss et al., 2002)

	Advantages	Disadvantages
Single cases	Greater depth	Limits on the generalisability of conclusions drawn. Biases such as misjudging the representativeness of a single event and exaggerating easily available data
Multiple cases	Augment external validity	Less depth per case

However, this single case design might have limitations (Leonard-Barton, 1990, Voss et al., 2002): a) single cases have limits on the generalisability of the conclusions because models or theories are developed from one case study; b) the limit of generalisability implies the risks of misjudging the representativeness of a single event, and of exaggerating easily available data. These risks are also present in multiple cases, although these are mitigated (Voss et al., 2002). However, as stated above, analytic generalisation in lieu of statistical generalisation can be used for single cases as well as multiple cases.

Yin (2003) postulated five rationales for single case designs as shown in Table 3-4. The case of spare parts in the South Korean Navy might represent a typical military logistical case. This case might represent an extreme case as well, for its extremely non-normal demand features. These rationales could serve as the main reasons for conducting this single case study.

Table 3-4 Five rationales for single case design (Yin, 2003)

a) When it represents the critical case in testing a well-formulated theory;
b) When the case represents an extreme case or a unique case;
c) A single case is the representative or typical case;
d) A single case study is the revelatory case; and
e) A single case study is the longitudinal case: studying the same single case at two or more different points in time.

Furthermore, Yin (2003) sub-categorised the single and multiple cases as a two × two matrix: single-case versus multiple-case × holistic versus embedded. A single case can involve more than one unit of analysis: a case study design involving embedded units is called an embedded case study design; a case study design examining only the global nature of an organisation is called a holistic design (Yin, 2003). The case of spare parts in the South Korean Navy can be considered to be a holistic single case design if the case is considered at the Naval supply centre of the Naval Logistics Command (NLC). However, the case can be also regarded as an embedded single case design if the case is examined at the depots of the Naval bases. There are four major Naval bases and four minor Naval bases in the South Korean Navy (Saunders, 2009). This research employed the holistic single case design, because the focus of this research was to evaluate forecasting performance rather than inventory system performance. Therefore, the complex multi-echelon inventory systems consisted of various suppliers, the Naval supply centre, several depots at Naval bases, and warships in the South Korean Navy were not required to be described in detail.

3.6 Evaluation of Research

Quantitative empirical case study research should be designed to test the validity of quantitative theoretical models and quantitative theoretical problem solutions, with respect to real-life operational situations (Bertrand and Fransoo, 2002). Yin (2003) discussed four tests relevant to evaluating the quality of a research design: construct validity, internal validity, external validity, and reliability.

3.6.1 Construct validity

Construct validity tests correct operational measures for the concepts being studied, and ensures consistency between theory and the defined construct (McCutcheon and Meredith, 1993, Yin, 2003). Bertrand and Fransoo (2002) argued that operational research studies generally lack construct validity because data could be affected by subjective judgements. However, the major data for this research, historical consumption of spare parts, were obtained from the logistical database in the Naval Logistics Command. Therefore, construct validity might not to be a problem. However, the non-normal demand features of spare parts cause distrust in the data generating process. This is discussed in Chapter 4.

3.6.2 Internal validity

Internal validity demonstrates a causal relationship, in which certain conditions lead to other conditions (Yin, 2003). Internal validity is used for explanatory (or causal) studies only (Yin, 2003). This research is to investigate the causal relationships between forecasting methods and the performance of the forecasting methods, and between demand features and the performance of the forecasting methods. This research employs various accuracy measures for establishing the causal relationships including derivative measures using simulation in Chapters 5 and 6.

In Chapter 6, a classification model which predicts the relative performance of alternative forecasting methods will be proposed. In order to establish internal validity for a predictive model, validation is especially important (Hosmer and Lemeshow, 2000). Model diagnostics refers to an assessment of the quality of the model that have been specified and estimated (Cryer and Chan, 2008, p. 8). An example of diagnostics is measuring the fit of the data to the model and the overall influence of an observation upon the model. This research employs various diagnostic tests for establishing internal validity of the classification model in Chapter 6.

Resubstitution refers to an estimate which uses a data set to build the model as well as to test the model (White and Liu, 1997). The resubstitution is likely to present an overly optimistic view of the true accuracy of the model (White and Liu, 1997). One common method for establishing internal validation is cross-validation. Cross-validation is defined as "assessing the accuracy of a model across different samples" (Field, 2009, p. 784). The procedure of cross-validation can be conducted as: a model is built using the first portion of the data (namely training set); then its ability to predict an outcome is evaluated using the second portion of the data (namely test set) (Altman and Royston, 2000).

When the sample size for cross-validation is small, accurate estimation of the internal validity of the predictive model is difficult (Altman and Royston, 2000, Steyerberg et al., 2001). This requires a careful choice of a ratio for splitting data. A variety of ratios for splitting data are possible. For example, 50% cross-validation uses 50% of the data to test a model developed on 50% of the sample, then this process repeats two times, so that all data are tested for the model (Steyerberg et al., 2001). Steyerberg et al. (2001) tested cross-validation for the logistic regression model which is the classification model with this research. They fitted the logistic regression model with 30-day mortality data set of 40,830 patients with acute myocardial infarction, and then compared the splitting ratios. They argued that the resubstitution presented the over-optimistic result and 10% cross-validation produced lower bias and less variable estimation, compared with 50% cross-validation. The bias is referred to as the difference between estimated performance (i.e. performance in the training set) and test performance (performance in the test set). The performance of 50% cross-validation is likely to underestimate the model because only half of the data are used to build the model; the performance of 50% cross-validation is likely to be more variable because half of the data are used for validation (Steyerberg et al., 2001). Therefore, this research also employs 10% cross-validation for establishing internal validity of the classification model in Chapter 6.

3.6.3 External validity

External validity tests whether a study's findings are generalisable beyond the immediate case study (Yin, 2003). External validity is a major barrier in conducting case studies (Yin, 2003). Internal validity can be seen as an approximation to external validity (Steyerberg et al., 2001). By establishing internal validity, the applicability of the model to other data sets in the South Korean Navy could be approximated.

As a single case study design, the case of the South Korean Navy can be used in generalising existing theory, because a case study relies on analytic generalisation (Yin, 2003). The case of the South Korean Navy can also be used in extrapolating the theory to other situations (e.g. other military forces or airline industry) relying on logical analysis.

3.6.4 Reliability

Reliability tests whether the operations of a study can be repeated with the same results (Yin, 2003). Reliability was maintained by revealing every data source and every reference explicitly, and presenting every equation and every process in models adopted in the research transparently, so that any calculations are able to be audited. In order to maintain the reliability of the forecasting performance, the forecasts generated were examined by a variety of measures.

3.7 Research Procedure

The research procedure of this research is described as Figure 3-1. As a model-based single case study, this research started from establishing the research questions in Chapter 1. As mentioned, research questions are crucial to establishing a well-defined focus at the start and to directing the data collection. This is followed by theory development.

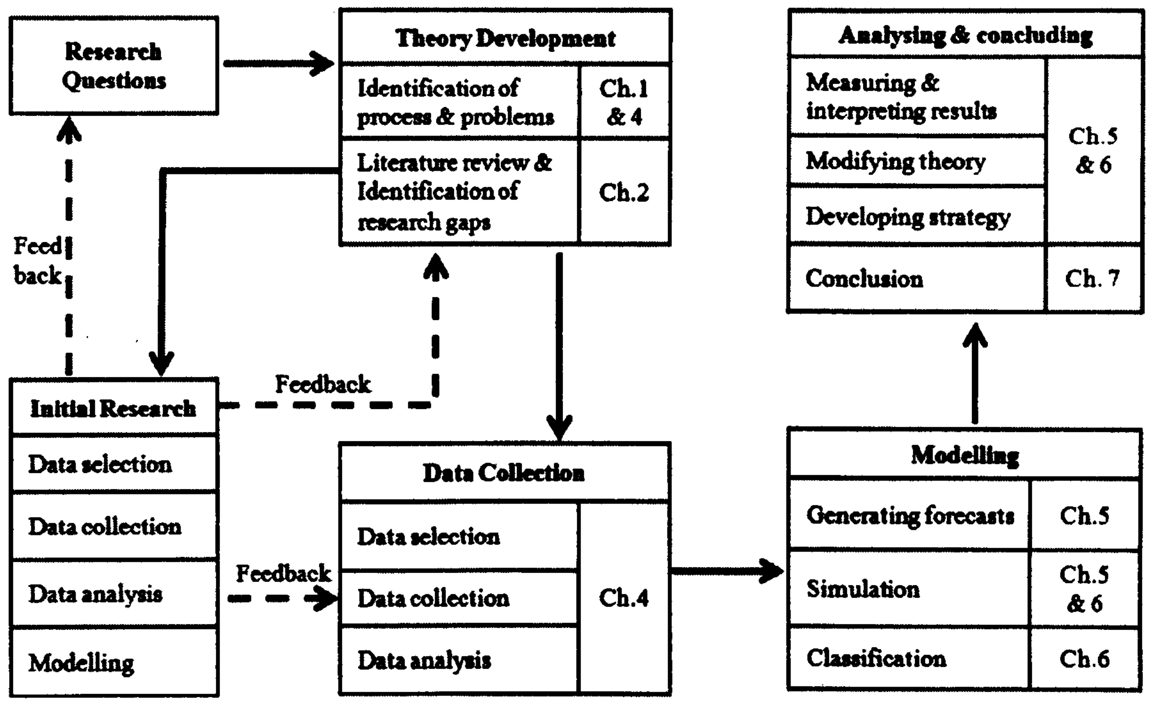


Figure 3-1 Research procedure

In the theory development stage, the forecasting process of the South Korean Navy and its problems were identified in Chapters 1 and 4. After reviewing literature relevant to these problems, the research gaps were identified in Chapter 2. Then, initial research which combines initial data selection, initial data collection, initial data analysis, and initial modelling was conducted. During the initial research period, a variety of analysis and modelling based upon the previous research, which was discussed in Chapter 2, were attempted. For example, lag-1 autocorrelations of the spare parts data obtained from the South Korean Navy were tested in the initial data analysis stage because lag-1 autocorrelation of the demand was identified to be a demand feature which influenced upon the performance of hierarchical forecasting (Widiarta et al., 2006). Businger and Read (1999) argued that the Box-Jenkins models can be useful in forecasting extremely volatile demand associated with many items of the US Naval spare parts. Thus, in the initial modelling stage, the Box-Jenkins models were attempted to be fitted to the data obtained. The initial research results provided feedback which requires a further literature review. This feedback clarified the research questions and guided further data collection. The stage of data collection is described in Chapter 4.

Following the data collection stage, the modelling stage was carried out based on the theory developed: the various competing forecasts are generated and measured in Chapter 5; simulation is conducted to verify forecasting performance in Chapters 5 and 6; and the classification model is proposed in Chapter 6.

In the stage of analysing and concluding, the results of empirical modelling are measured and interpreted in view of forecasting accuracy as well as the inventory context in Chapters 5 and 6. Depending on the results and the existing theory, the theory is modified and generalised as analytic generalisation. Also, the theory contributes to the development of a forecasting strategy for the South Korean Navy. Then, concluding remarks are presented in Chapter 7.

3.8 Summary and Conclusion

This chapter presented a methodological framework for this research. The purpose of this research might be in line with an exploratory study as well as an explanatory study. As an operations management study, this research might be empirical normative quantitative research and empirical quantitative case study.

Since a case study is suitable for "how" and "what" questions as well as exploratory research and explanatory research, which are the cases with this research, this research employed a case study as the research strategy.

Theories relevant to this research are available from the existing theories, although the necessary theories are not well-developed for the context of this research. However, theories from similar contexts can be the theoretical foundation of this research. The theory of this research can be formed by an inductive method and can generalise as analytic generalisation.

Research design is the logical sequence that links the empirical data to research questions (Yin, 2003). The four components of research design were reviewed: research questions; the unit of analysis; the logic linking the data to the research questions; and the criteria for interpreting the findings. This research employed the single case research design because this provides more opportunities for in-depth observation, and the single case design is appropriate for extreme cases and typical cases which are the case involved with this research. Then, four tests which can evaluate the research were reviewed. Finally, the overall research procedure was described.

Based on the research topic presented in Chapter 1, the related literature was reviewed in Chapter 2, and the research design of this research was presented in this chapter. This thesis moves into the empirical case from the next chapter. The finding chapters start with the analysis of the data, which is the nature of the spare parts demand in Chapter 4.

Chapter 4. Nature Of The Spare Parts Demand

As stated earlier, the core of the problem for forecasting the Naval spare parts demand might be the non-normality of the demand and the inappropriate forecasting methods of the Navy. The non-normality of the spare parts demand in the South Korean Navy was expected as explained in Subsection 1.2.2. The impact of demand features upon forecasting performance was reviewed in Section 2.5. The identification of the nature of spare parts demand of the South Korean Navy is important as a prerequisite for the investigation into forecasting methods. This chapter analyses the nature of the spare parts demand in the South Korean Navy.

This chapter starts by presenting the general information about the fleet of the South Korean Navy and the spare parts demand of the Navy in Section 4.1. This is followed by reviewing the inventory control methods in the South Korean Navy in Section 4.2. Clarification of the general information and the inventory control methods of the South Korean Navy might be useful to understand the demand features of the spare parts. The inventory system of the South Korean Navy will also be the framework for the simulation experiment to measure forecasting performance in the next chapter. Section 4.3 relates the demand features of the spare parts. The spare parts demand is analysed with various techniques including the various statistics which were reviewed in Section 2.5. In Subsection 1.2.2, it was stated that the difficulty of forecasting spare parts demand might arise from the non-normality of the spare parts demand data. In Subsection 2.5.2, simple exponential smoothing has been reported as a superior forecasting method for hierarchical forecasting strategy. In order to use simple exponential smoothing trend and seasonal components need to be removed or measured.

Section 4.4 presents the procedure for removing the trend and seasonal components in the time series of the Naval spare parts. In Chapter 3, it was stated that the non-normal demand features cause distrust in the data generating process with respect to construct validity. The sources of non-normality in the demand are reviewed in Section 4.5. Finally, a summary and concluding remarks are presented in Section 4.6.

4.1 General Information

The South Korean Navy consists of 184 warships including 22 submarines, 10 destroyers, 9 frigates, 28 corvettes, and 82 fast attack crafts (Saunders, 2009). Warships are characterised as having high construction costs. For example, the construction costs for a destroyer with Aegis combat control systems of the South Korean Navy are known to be ₩1.2 trillion (£613 million)[1] (Defence Industry Daily, 2009).

There are four major Naval bases in the South Korean Navy: fleet headquarters in Chinhae, the 1st fleet in Donghae, the 2nd fleet in Pyongtaek, and the 3rd fleet in Pusan (Saunders, 2009). There are also four minor Naval bases in Cheju, Mokpo, Mukho, and Pohang (Saunders, 2009). In order to supply spare parts to warships, complex multi-echelon inventory systems consisting of various suppliers, the Naval supply centre, several depots at the major and minor Naval bases, and warships have been established.

A large stock of spare parts is held for the 184 warships in the Navy. However, there is no demand for a large proportion of the stocked items. While 45,557 warship spare parts items were held, the demand in 2008 was for 26,415 of these (Seon and U, 2009). Table

---

[1] 1.00 British pounds sterling (£) = 1,957.00 South Korean Won (₩) on 22nd May 2009 (www.currencyconventor.uk.com)

4-1 presents the classification of spare parts items for warships by value or demand volume in the South Korean Navy in 2008. A large proportion of spare parts are low value or low demand items. 62.13% of spare parts were less than ₩100,000 (£51.1) in terms of unit cost. 52.3% of spare parts had a demand of one or zero.

Table 4-1 Classification of spare parts (Seon and U, 2009)

Classification by value				Classification by demand volume			
Value	Unit cost	No. of items	%	Volume	Annual demand	No. of items	%
Low	0 ~ ₩100,000	28,303	62.1	Low	0 ~ 1	23,838	52.3
Medium	₩100,001 ~ ₩5,000,000	14,341	31.5	Medium	2 ~ 5	11,159	24.5
High	₩5,000,000 ~	2,913	6.4	High	6 ~	10,560	23.2
Total		45,557	100	Total		45,557	100

1.00 British pounds sterling (£) = 1,957.00 South Korean Won (₩) on 22nd May 2009 (www.currencyconventor.uk.com).

Although the unit cost of most of the spare parts is low, the stock-out costs can dramatically outweigh the unit cost. For example, the absence of a £10,000 spare part might cause a £100 million warship to be non-operational. This could even lead to a military defeat that could cause casualties and deaths (MacDonald, 1997).

Figure 4-1 presents time plots of annual mean demand per item in each value group. Annual demand volumes for low value items were significantly higher than high and medium value items. Whilst similar volumes of high and medium value items were demanded throughout the whole period, higher volumes of low value items were demanded in 2002 ~ 2003 than in 2004 ~ 2007. As such, low value items are characterised with downward trends throughout the period.

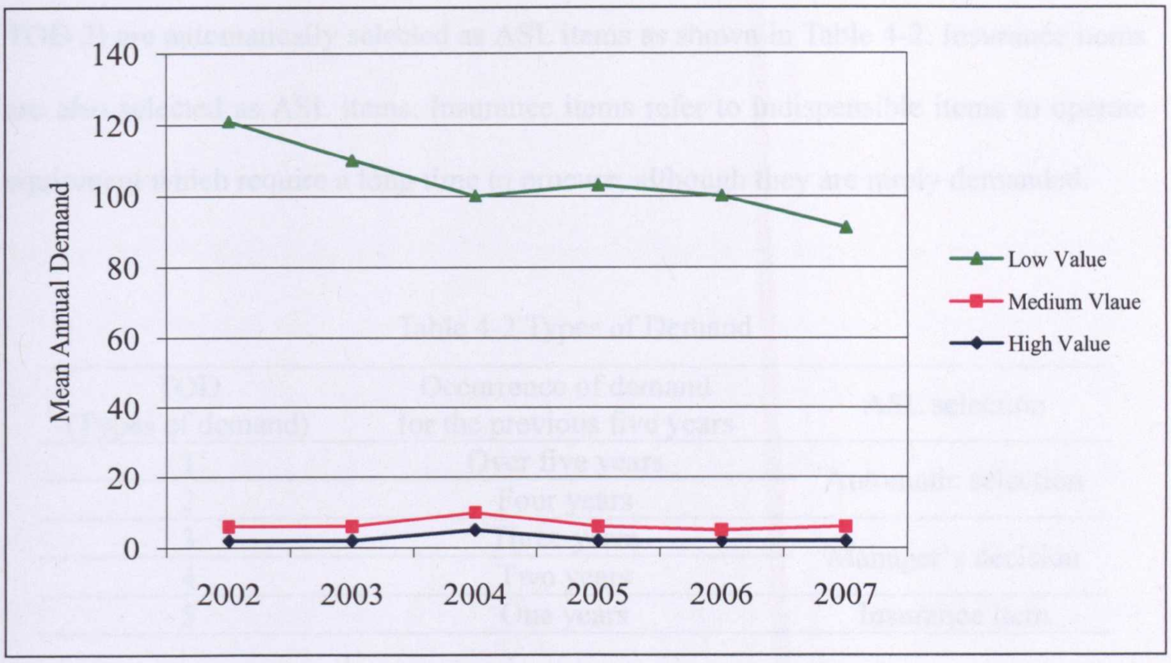


Figure 4-1 Mean annual demand for items in three value groups (Seon and U, 2009)

4.2 Inventory Control Method

The inventory control methods for spare parts in the South Korean Navy are based on Authorized Stock Item List (ASL). ASL is the list of items which are sanctioned to be stocked within each depot and the Naval supply centre; however, items which are not in the list [i.e. Non-Authorized Stock Item List (N-ASL)] are not required to be stocked. This section describes the inventory control methods for the ASL and N-ASL items.

4.2.1 Authorised Stock Item List

The South Korean Navy differentiates the inventory control methods for ASL items and N-ASL items. An order of an N-ASL item is placed only when a backorder of the item is placed. Items with at least one demand occurrence in two of the previous five years (i.e. TOD 1 ~ TOD 4) are considered to be selected as an ASL item; among them, items with at least one demand occurrence in four of the previous five years (i.e. TOD 1 and

TOD 2) are automatically selected as ASL items as shown in Table 4-2. Insurance items are also selected as ASL items. Insurance items refer to indispensible items to operate equipment which require a long time to procure, although they are rarely demanded.

Table 4-2 Types of Demand

TOD (Types of demand)	Occurrence of demand for the previous five years	ASL selection
1	Over five years	Automatic selection
2	Four years	
3	Three years	Manager's decision
4	Two years	
5	One years	Insurance item

4.2.2 Inventory objective of Authorised Stock Item List items

The inventory systems for warship spare parts in the South Korean Navy consist of four interconnected echelons in which spare parts flowing through the systems are stored at different points. These four echelons are various suppliers, the Naval supply centre, 8 depots, and 184 customers (i.e. warships). Such inventory systems are known as multi-echelon inventory systems (Slack et al., 2004). The South Korean Navy uses the terms retail-level and wholesale-level. The retail-level refers to the echelon with the depots; the wholesale-level refers to the echelon with the Naval supply centre. Each major or minor Naval base has its own retail-level depots. Each warship also carries a limited range of spare parts, so as to deal with routine maintenance or minor breakdown at sea. In principle, spare parts for warships are supplied from retail-level (i.e. the depots); and spare parts stored at retail-level are supplied from wholesale-level (i.e. the Naval supply centre).

As inventory objective refers to an order-up-to-level either at wholesale-level or at retail-level. The South Korean Navy distinguishes the inventory objectives of Authorized Stock Item List items between wholesale-level and the retail-level. The Requirements Objective (RO1) indicates the inventory objective at the wholesale-level and the Requisition Objective (RO2) denotes the inventory objective at the retail-level. Table 4-3 describes the corresponding composition of the inventory objectives at each level. PC indicates the expected quantity of demand during the intervals of procurement at the wholesale-level as a similar concept to OL at the retail-level. SL means the safe stock level. PROLT means the real empirical lead time at the wholesale-level including lead times for administration, manufacturing and delivery. OST indicates the lead time at the retail-level including order and shipment times.

Table 4-3 Inventory objective of the South Korean Navy

Wholesale-level Requirements Objective (RO1)	Retail-level Requisition Objective (RO2)
Procurement Cycle Quantity (PC)	Operating Level (OL)
Safety Level (SL)	Safety Level (SL)
Procurement Lead Time (PROLT)	Order & Shipping Time (OST)

The South Korean Navy differentiates the inventory control methods for the retail-level and the wholesale-level. The inventory control method of retail units is similar to the (s, S) control system; that is, continuous review, order-point, order-up-to-level system (Silver et al., 1998). As shown in Subsection 2.7.2, inventory position can be calculated as "inventory position = (stock on hand) + (stock on order) – (backorders)" (Silver et al., 1998). When the inventory position of a retail unit is below a reorder point, s (s = SL + OST), the unit places an order to raise the inventory position to RO2 automatically. On the other hand, the inventory control method of wholesale units is similar to the (R, S) control system; that is, the periodic-review, order-up-to-level system (Silver et al., 1998). Although the enacted review cycle (i.e. procurement decision cycle) is one year, the modification of the orders is allowed for an exceptional case throughout a year. Generally, the wholesale unit places an order to raise inventory position to RO1 (RO1 = PC + SL + PROLT) once a year, in line with legal requirements.

An order of an ASL item can come from two sources. For the purpose of repairs and maintenances of warships in the Naval shipyard of the Naval Logistics Command (NLC), at the repair shops of Naval bases, and on the warships themselves, customers (i.e. warships) place an order to a retail unit. For the purpose of keeping their inventory positions, retail and wholesale units also place an order to wholesale units and suppliers respectively.

This section identified that the inventory control method for the spare parts in the South Korean Navy is based upon the Authorized Stock Item List. The inventory objectives in both retail- and wholesale-levels, and their control methods were also clarified. Upon the basis of this information about the inventory control methods together with the general information described in Section 4.1, the demand features of spare parts which are the main concerns of this chapter are analysed as follows.

4.3 Spare Parts Demand Analysis

In Subsection 1.2.2, it was mentioned that the spare parts demand in the South Korean Navy might be non-normal. It was also mentioned that hierarchical forecasting strategy is based upon a hierarchical demand structure. In Section 2.5, it was reviewed that the demand features such as correlation and forecasting horizon influence forecasting performance. In this section, the spare parts demand features in the South Korean Navy are analysed. This section starts by presenting the structure of National Stock Number

(NSN) on which the Naval spare parts are based. The processes of data selection and grouping then follow. Then, detailed description of the data with respect to the techniques reviewed in Chapter 2 is presented.

4.3.1 National Stock Number

National Stock Number is a thirteen-digit numeric code used by North Atlantic Treaty Organisation (NATO) to classify standardised material items of supply as shown in Figure 4-2.

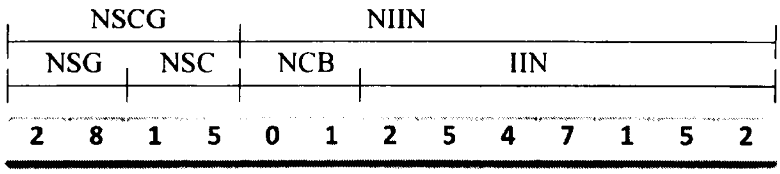


Figure 4-2 National Stock Number (NSN)

The first two-digits, NATO Supply Group (NSG), classify items according to their end use (see Appendix B); then, the next two-digits, NATO Supply Class (NSC), sub-classify them according to their shape and size. The last nine-digits represent National Item Identification Number (NIIN). The first two-digits of NIIN, National Codification Bureau (NCB), indicate the country of origin (e.g. 01 is assigned to the USA; and 99 is assigned to the United Kingdom). The next seven-digits, Item Identification Number (IIN), are simply allocated without a pattern. In Figure 4-2, the first two-digits, "28" indicates NSG 28 (i.e. engines, turbines and components); the first four-digits, "2815" denotes NSCG 2815 (i.e. diesel engines and components); the next two-digits, "01" indicates NCB 01 (i.e. the USA); and the last seven-digits, "547152" denotes a serial number which has no specific pattern. Therefore, the exemplified thirteen-digit code, "2815012547152" represents a part (i.e. piston ring) of a generator (i.e. MTU 6V396). As such, National Stock Number represents a hierarchical demand structure. The South Korean Navy utilises National Stock Number.

4.3.2 Data selection and grouping

Before presenting the demand features, it is necessary to clarify the process of data selection and grouping methods required to generate hierarchical forecasting. Generally, the ratio of the ASL to all spare parts in the South Korean Navy is merely eleven percent; meanwhile, approximately eighty percent of the total request is replenished by ASL items. As mentioned in Subsection 2.7.1, by the rule of 80/20, managing high frequency items might be crucial in increasing spare parts supply within budgetary limitations so as to maximise operational availability. Unlike business data, the range of frequency in the South Korean Navy is relatively low. High frequency items in militaries can be considered to be low frequency items in business (Eaves and Kingsman, 2004). For example, Sani and Kingsman (1997) classified 'annual demand less than 20 units' into 'very low demand' for 30 daily spare parts demand data for vehicles and agricultural machinery over five years as shown in Table 2-9.

Although TOD 1 and TOD 2 of Table 4-2 are still of very low demand in the classification scheme of Sani and Kingsman (1997), TOD 1 and TOD 2 are considered to be high frequency items for the South Korean Navy. The ratio of TOD 1 and TOD 2 to all spare parts is merely four percent, however, TOD 1 and TOD 2 account for approximately sixty percent of the total requests. Managing TOD 1 and TOD 2 might be crucial to increase the operational availability of weapon systems. Low frequency items (lower than TOD 2) are extremely difficult to analyse. Hence, low frequency items were screened out in this research. Managing low frequency items (lower than TOD 2) is beyond the scope of this research. These might be able to be managed with a strategic collaboration between suppliers and users such as military supply chain integration (Chen et al., 2005).

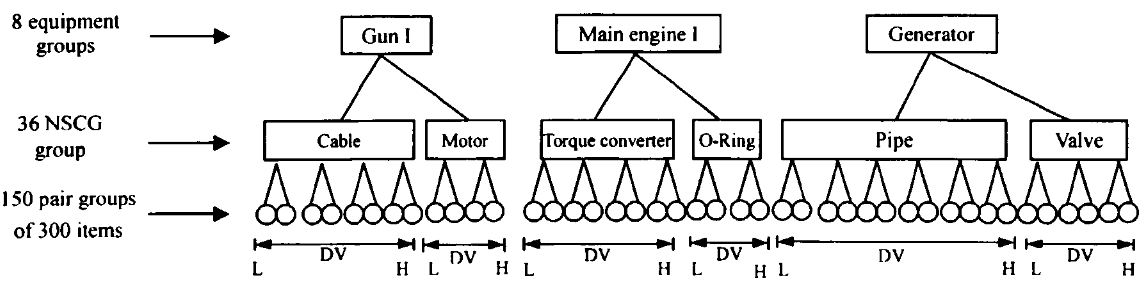
Historical demand records (2002 ~ 2007) for three types of warships (type A, B and C) were collected for this research. As stated in Subsection 3.5.1, a specific group of warships and a time boundary were clarified. The time boundary was decided to be from January 2002 to November 2007 because the Naval maintenance data system, which is the major data source, has been stabilised since 2002. Among the demand records, those for eight pieces of equipment installed in the three kinds of warships were selected. These eight pieces of equipment are important for sustaining the operational availability of those warships. Table 4-4 presents the description of the selected data.

Table 4-4 Selection of data (300 items)

Equipment	No. of items			Type of ship
	Selected	TOD 1 - 2	Total	
Gun I	22	27	155	C
Gun II	10	12	760	A, B, C
Gun III	6	7	1,038	A, B
Main Engine (ME) I	54	65	1,830	A, B
Main Engine (ME) II	134	161	2,489	C
Generator (GE) I	56	67	2,036	A, B
Air Compressor (AC) I	12	15	595	A, B, C
Radar (RD) I	6	7	466	B, C
Total	300	361	9,369	

Overall, 9,369 items were identified which met the above conditions. Among the 9,369 items, 361 items were identified to be either TOD 1 or TOD 2. For the purpose of grouping in pairs, and for the ease of analysing the data and understanding the results, an even number of samples, rounded to the nearest hundred (i.e. 300) were chosen from the 361 items. An even number sample from each equipment group was chosen using a random sampling procedure with a random number generator (Howitt and Cramer, 2008). The dataset therefore comprised 300 items, which were selected for inclusion in forecasting models.

In order to use a hierarchical forecasting method, the form of a group needs to be defined. An exemplified grouping structure for this research is presented as in Figure 4-3. Grouping is based on the types of equipment and National Stock Number (NSN). The 300 items were classified by the 8 types of equipment which the items are used for. Then, the items of each type of equipment were sub-classified into 36 groups using the NATO Supply Classification Group (NSCG). Intuitively, as shown in Table 1-3, the hierarchical assembly structure of Naval spare parts consisted of parts and components, and so an assembly structure might capture a more practical hierarchical demand structure than the NSCG. However, the data required to identify the assembly structure such as assembly drawings and data on failure of assemblies are classified and hence not available for this research.



Key: DV = historical dollar volume (historical demand per year × item unit price); L = low DV; H = high DV.

Figure 4-3 Grouping structure

As stated in Subsection 2.5.2, Fliedner and Mabert (1992) have claimed that grouping criterion based on historical dollar volume (DV) provides significantly better performance for hierarchical forecasting. Hence, the items were ranked in terms of DV (defined as historical demand per year × item unit price) within the same type of equipment and the same NSCG.

As mentioned in Subsection 2.5.2, research has found that the number of items in a group (or the number of group with a limited number of items) has no significant effect upon the performance of hierarchical forecasting (Fliedner and Mabert, 1992, Fliedner and Lawrence, 1995). Fliedner (1999) noted that the different numbers of items in the groups is a source of incongruity of the performance of hierarchical and direct forecasting methods. Much of the research about the influence of correlations upon the performance of hierarchical and direct forecasting methods limited the number of items in a group to two items (Schwarzkopf et al., 1988, Dangerfield and Morris, 1992, Fliedner, 1999, Widiarta et al., 2006, Widiarta et al., 2008a, Widiarta et al., 2009). In order to test the impact of correlations on forecasting performance, avoid any biasing effect which could be introduced by complex interrelationships among more than two items in a group, and compare the results of this research with the previous research, the experimental design of two item groups was employed for this research.

Grouping an item with another similar item can bring out hidden patterns and decrease errors (DeLurgio, 1998). Hence, the two nearest items (i.e. the most homogeneous) in terms of historical dollar volume (DV) formed a group. The 300 items can be considered as 150 pair groups for hierarchical forecasting, each containing two items.

4.3.3 Measures of demand features

A variety of measures to identify demand features were reviewed in Section 2.5. This research employs measures from the literature. Before analysing the spare parts demand data, this subsection clarifies the measures and their properties to be used.

In Subsection 2.5.2, lag-1 autocorrelation of a time series was discussed as a measure to identify demand feature which influences upon the relative performance of top-down and direct forecasting (Widiarta et al., 2006). If the lag-1 autocorrelation, $\rho(1)$, of the time series for at least one of the items in a group consisting of two items is greater than 1/3, direct forecasting outperforms top-down forecasting. However, if the lag-1 autocorrelations of the two item level time series are satisfied, $-1 < \rho(1) \leq 1/3$, the difference in the performance of the two forecasting strategies is non-significant. In the initial data analysis stage of this research, the lag-1 autocorrelations of 272 monthly, 252 quarterly, and 286 yearly time series of the 300 item level time series were found to be in the scope of non-significant autocorrelations ($-1 < \rho(1) \leq 1/3$). Thus, the autocorrelation was not considered to be a measure used to identify a data feature in this research. The spare parts demand data were analysed by the following measures.

Identification of trend and seasonal components is important because these are predictable components of a time series (Bowersox and Closs, 1996, Silver et al., 1998). As stated in Section 2.1, Ghobbar and Friend (2002, 2003) examined the time series of aircraft spare parts demand and identified trend, seasonal and irregular components. Businger and Read (1999) used trend and seasonality to identify demand features of spare parts in the US Navy. In this research, Slope denotes the gradient of a linear regression model fitted to the time series data. The processes and the results of tests for seasonal effects will be presented later in this section.

As stated in Subsection 2.5.1, coefficient of variation in demand size is a unit free measure of relative variability. Coefficient of variation in demand size was used as a classification criterion for non-normal demand in previous literature (Williams, 1984, Businger and Read, 1999, Syntetos, 2001). In this research, Cv(size) indicates the coefficient of variation in demand size. Equation (2-32) was employed to express Cv(size). As stated in Subsection 2.5.1, the erratic and lumpy demand features in the categories of non-normal demand in Subsection 1.2.2 might be captured by this measure.

The number of periods with zero demand is a statistic for measuring the intermittency of demand (Businger and Read, 1999, Boylan et al., 2008). As stated in Subsection 2.5.1, demand features reflecting intermittency (i.e. the intermittent, slow moving, clumped and lumpy demand features) in the categories of non-normal demand might be captured by this statistic. A problem with this statistic was identified in Subsection 2.5.1. It was noted that the number of zero demand periods depends on the overall data periods; that is, the number of zero demand periods might be longer when the overall data periods are longer, and vice versa. Boylan et al. (2008) employed the number of demand periods with zero demand during the last n time periods. They used the last 13 time periods (n = 13) for their classification. However, they offered no rationale for this decision about the time periods, n.

In this research, different data periods were utilised to generate different forecasts in different years. For example, in order to generate a monthly forecast in 2005, 36 monthly data points (i.e. between January of 2002 and December of 2004) were utilised; in order to generate a monthly forecast in 2006, 48 monthly data points (i.e. between January of 2002 and December of 2005) were utilised. In order to compare all the forecasting results, a unit free measure for the number of zero periods was required. A unit free measure, the proportion of zero demand periods, namely Pr(zero), employed in this research is defined as:

$$\textit{Proportion of zero demand periods} = \frac{\sum_{t=1}^{n} I(y_t)}{n} \tag{4-1}$$

where:

y_t = the real demand size for an item at time period t

if $y_t = 0$, $I(y_t) = 1$; otherwise, $I(y_t) = 0$

n = the number of overall time series periods

The distribution of data can deviate in two ways from a normal distribution: the distribution can be skewed, when one tail of the distribution is longer than the other tail; and the distribution can be kurtosed, when the distribution is too flat or highly peaked (i.e. the tails are too long or too short) (Miles and Shevlin, 2001).

The lack of symmetry of a distribution around its mean can be identified by skewness (Tabachnick and Fidell, 2007). As mentioned in Subsection 2.5.1, Businger and Read (1999) used a simple formula of skewness to identify a demand feature of spare parts. Skewness for this research is defined as in equation (4-2). The skewness of a variable that is normally distributed has the value 0. Positive skewness denotes a distribution with an asymmetric tail stretching toward more positive values (right longer tail) – the mean and the median are larger than the mode; whereas negative skewness represents a distribution with an asymmetric tail clustered at the more negative values (left longer tail) – the mean and the median are smaller than the mode (Howitt and Cramer, 2008).

$$Skewness = \frac{n}{(n-1)(n-2)} \sum_{t=1}^{n} \left(\frac{y_t - \bar{y}}{s} \right)^3 \tag{4-2}$$

where:

y_t = the real demand size for an item at time period t

$\bar{y}$ = the mean demand size

s = the standard deviation of demand size

n = the number of overall time series periods

Kurtosis (pointyness) identifies the relative peakedness or flatness of a distribution compared with a normal distribution (Tabachnick and Fidell, 2007). The kurtosis of a variable that is normally distributed has the value 0. Positive kurtosis (i.e. leptokurtic distribution) represents a relatively pointy distribution; negative kurtosis (i.e. platykurtic distribution) represents a relatively flat distribution (Howitt and Cramer, 2008). Kurtosis for this research is defined as in equation (4-3).

$$Kurtosis = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{t=1}^{n} \left(\frac{y_t - \bar{y}}{s} \right)^4 \right] - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{4-3}$$

As stated in Subsection 2.5.2, some authors (Schwarzkopf et al., 1988, Viswanathan et al., 2008) found that the variability of demand size significantly influenced upon the relative forecasting performance of top-down and direct forecasting. In addition to Cv(size), this research used another measure for the variability of demand size reflecting the peak demand. As stated in Subsection 2.5.1, Businger and Read (1999) used the number of peaks as in equation (2-34). However, a unit free measure is required for this research because different data periods were utilised to generate different forecasts in different years. As a unit free measure, proportion of peak demands [Pr(peak)] is calibrated as in equation (4-4). In order to capture the variability of demand size, the sum of the peak demands was divided by the sum of the total demands. As stated in Subsection 2.5.1, the erratic and lumpy demand features in the categories of non-normal demand might be captured by this statistic.

$$\textit{Proportion of peak demands} = \frac{\sum_{t=1}^{n} I(d_t > 2) \times y_t}{\sum_{t=1}^{n} y_t} \tag{4-4}$$

where: $d_t = \dfrac{|y_t - \bar{y}|}{s}$

$$I(d_t > 2) = \begin{cases} 0, & d_t \leq 2 \\ 1, & d_t > 2 \end{cases}$$

As discussed in Subsection 2.5.2, the influence of correlations upon the relative performance of top-down forecasting and direct forecasting is the most controversial issue (Schwarzkopf et al., 1988, Gross and Sohl, 1990, Dangerfield and Morris, 1992, Fliedner, 1999, Widiarta et al., 2006, Widiarta et al., 2008a, Widiarta et al., 2009). In addition to the correlations of the item level time series with other item level time series in the same group [i.e. Corr(item)], the correlations of the item level time series with the group level time series [i.e. Corr(group)] composed of two item level time series were also examined. This was because hierarchical forecasting is expected to be influenced by the correlations between the group level time series and the item level time series [Corr(group)] through a proration method. A high Corr(group) is expected, because group level time series is a linear combination of item level time series.

As stated in Subsection 2.5.1, lead time variation was argued to be an important demand feature in the demand classification for direct forecasting (Williams, 1984, Eaves, 2002, Eaves and Kingsman, 2004). A long procurement lead time (PROLT) requires a long forecasting horizon for a procurement decision. As stated in Subsection 2.5.2, Shlifer and Wolff (1979) contended that forecasting horizon (reflecting PROLT) influences upon the relative performance of top-down and direct forecasting. Thus, this research examined PROLTs of the spare parts.

As shown in Subsection 2.5.2, UV (historical unit volume) and DV (historical dollar volume) were contended to be significant grouping criteria which improved the performance of top-down forecasting (Fliedner and Mabert, 1992). In this research, UV was calculated by monthly, quarterly or yearly mean demand (i.e. mean) for an item. DV was calculated as "DV = historical demand for an item per year × item unit price".

4.3.4 Historical demand analysis

Using the above mentioned measures, the spare parts demand data were analysed. This subsection presents the results of analysis. Data aggregation refers to the bucketing of individual data observations into time periods (Eaves, 2002). The time series plot of the sum of the 300 monthly aggregated spare parts demand time series is presented in Figure 4-4. There were two peak points in June 2002 and March 2003 followed by a downward trend. These two peak points could be explained by an increased operational demand on the warships caused by the sea battles with the North Korean Navy in June of 2002 (Jie-Ae, 2002) and subsequent preparation against possible clashes before the fishing season in the next year 2003. The downward trend of demand is consistent with the mean annual demand for total spare parts as shown in Figure 4-1. Although it is not obvious, there was a more or less seasonal fluctuation: high in December - March; low in May - October.

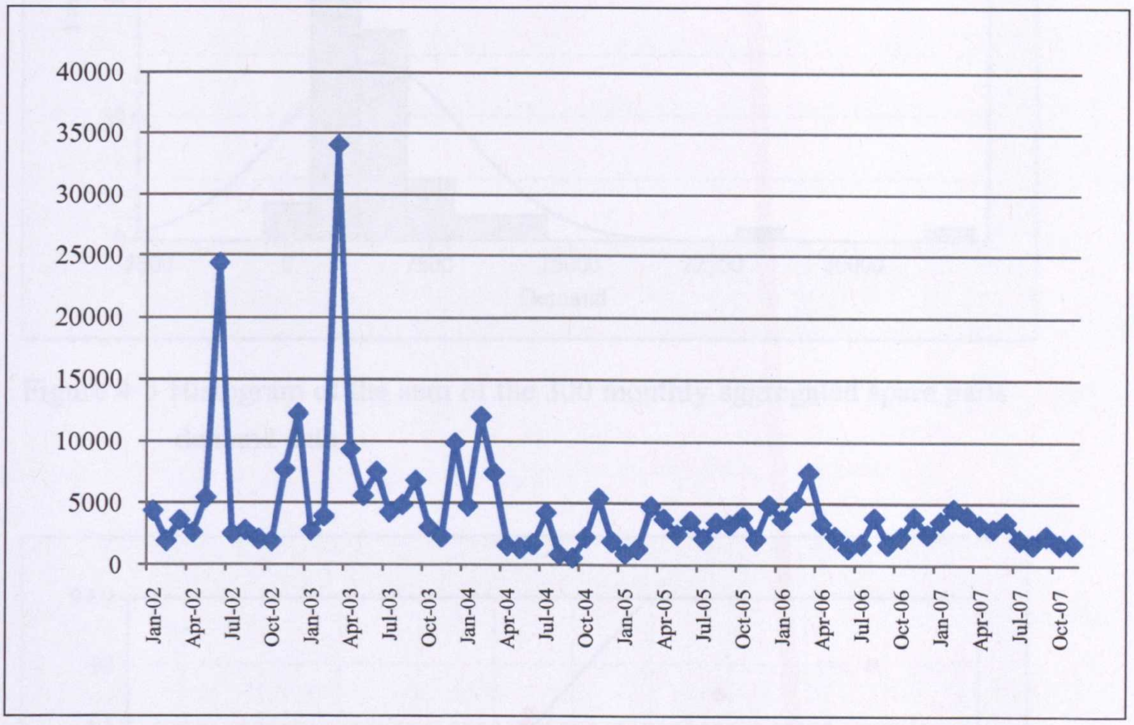


Figure 4-4 Time series plot of the sum of the 300 monthly aggregated spare parts demand time series

Figure 4-5 presents the histogram of the sum of demand quantities in the 300 monthly aggregated data. The data were significantly skewed toward the left and were highly peaked (i.e. leptokurtic distribution) as shown in Figure 4-5. The two outliers in June 2002 and March 2003 were identified as two separate columns.

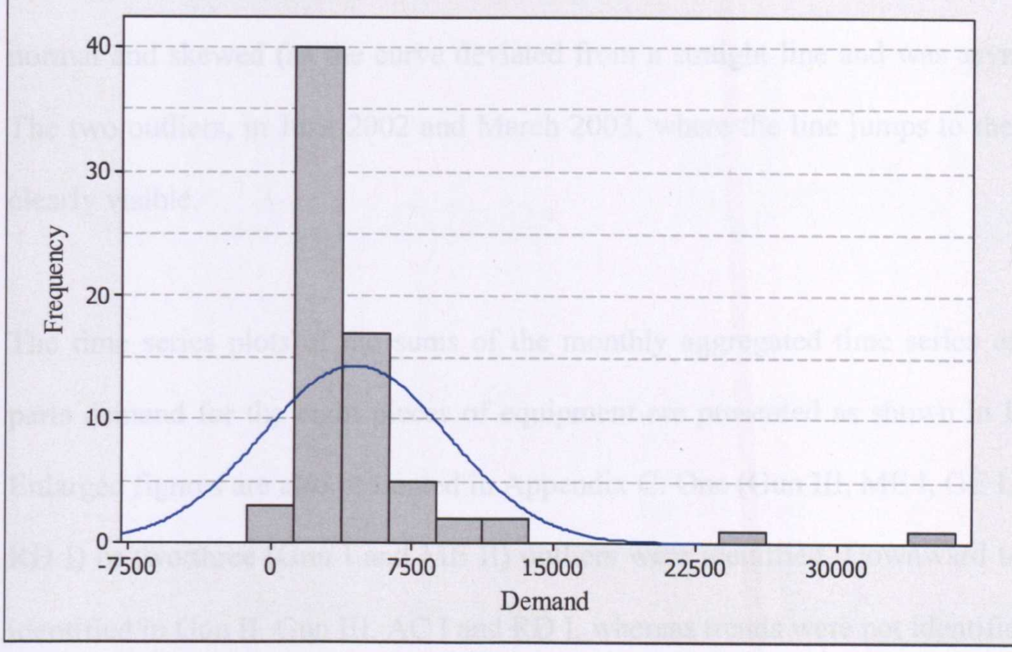


Figure 4-5 Histogram of the sum of the 300 monthly aggregated spare parts demand data

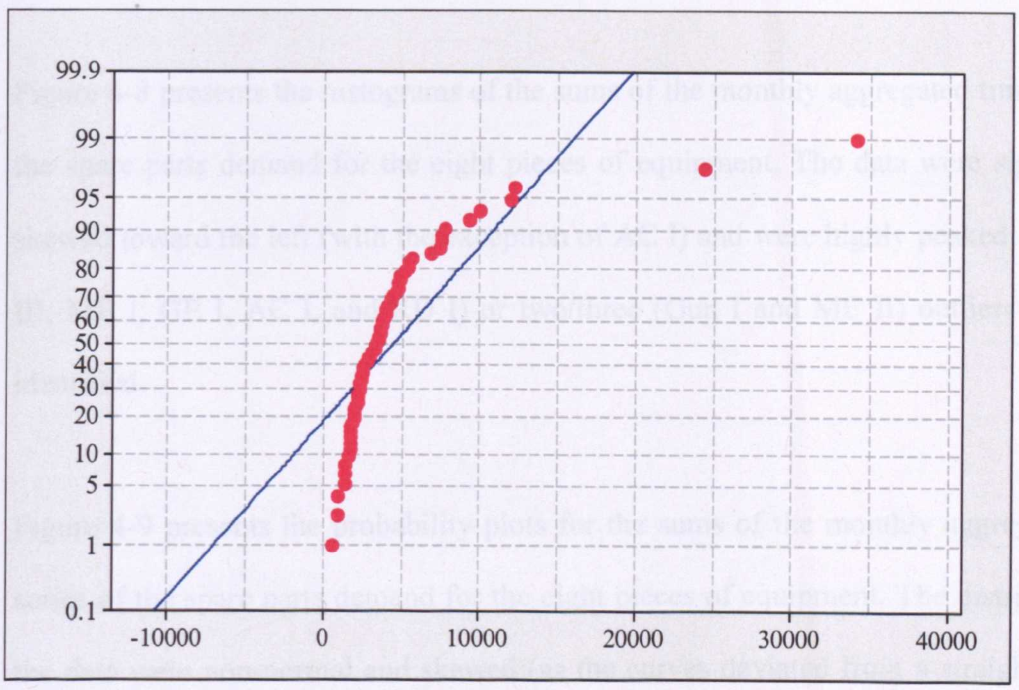


Figure 4-6 Probability plot of the sum of the 300 monthly aggregated spare parts demand data

Figure 4-6 presents the probability plot against a normal distribution for the sum of the

300 monthly aggregated spare parts demand data. The distribution of the data was non-normal and skewed (as the curve deviated from a straight line and was asymmetrical). The two outliers, in June 2002 and March 2003, where the line jumps to the end, were clearly visible.

The time series plots of the sums of the monthly aggregated time series of the spare parts demand for the eight pieces of equipment are presented as shown in Figure 4-7. Enlarged figures are also presented in Appendix C. One (Gun III, ME I, GE I, AC I, and RD I) or two/three (Gun I and ME II) outliers were identified. Downward trends were identified in Gun II, Gun III, AC I and RD I, whereas trends were not identified in Gun I, ME I, ME II, and GE I. Although lower demands for Gun I and AC I were identified in June and July, seasonality was not obvious in most of the time series.

Figure 4-8 presents the histograms of the sums of the monthly aggregated time series of the spare parts demand for the eight pieces of equipment. The data were significantly skewed toward the left (with the exception of AC I) and were highly peaked. One (Gun III, ME I, GE I, AC I, and RD I) or two/three (Gun I and ME II) outliers were also identified.

Figure 4-9 presents the probability plots for the sums of the monthly aggregated time series of the spare parts demand for the eight pieces of equipment. The distributions of the data were non-normal and skewed (as the curves deviated from a straight line and were asymmetrical). One (Gun III, ME I, GE I, AC I, and RD I) or two/three (Gun I and ME II) outliers, where the lines jump to the ends, were also identified.

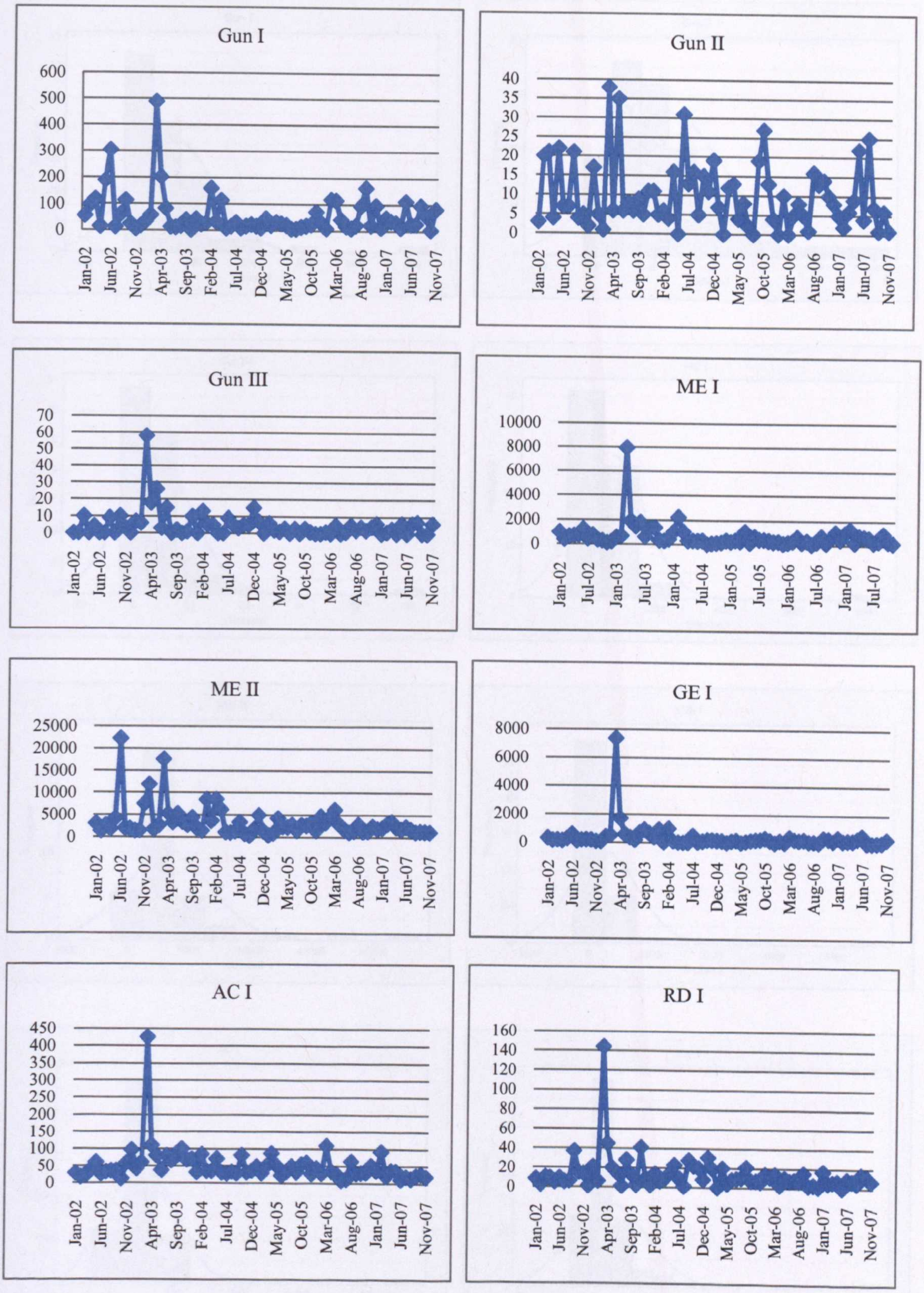


Figure 4-7 Time series plots of the sums of the monthly aggregated time series of the spare parts demand for the eight pieces of equipment

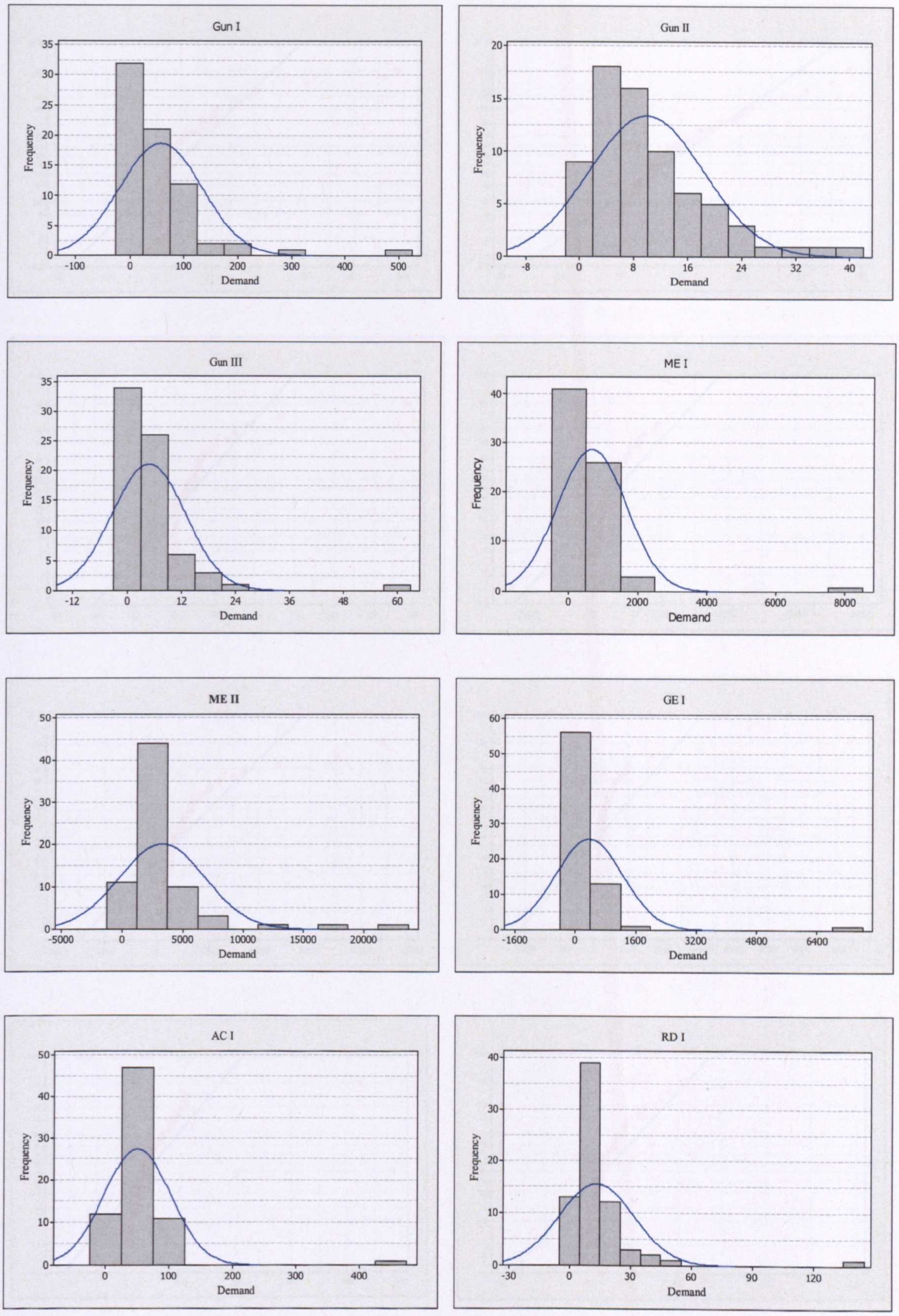


Figure 4-8 Histograms of the sums of the monthly aggregated time series of the spare parts demand for the eight pieces of equipment

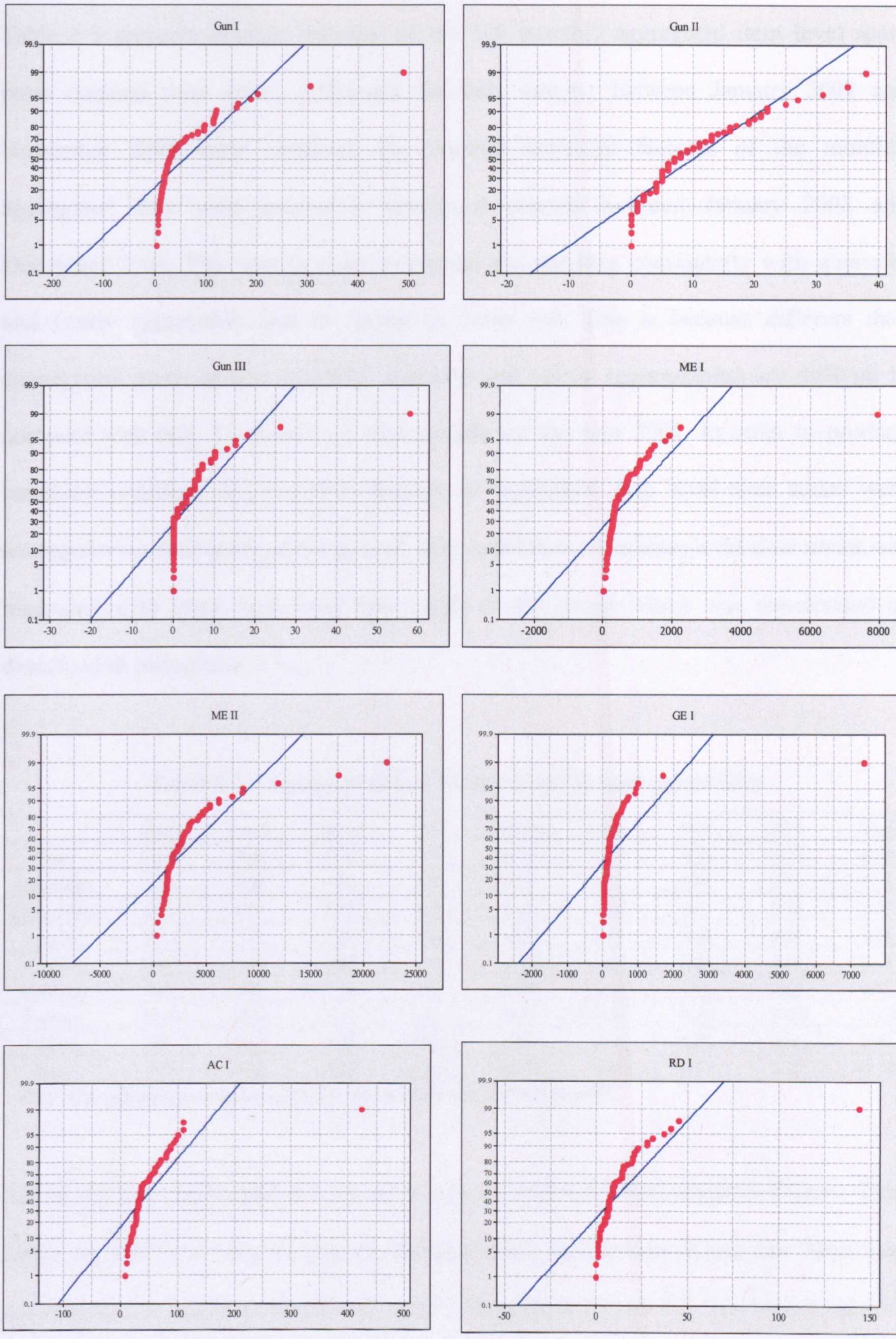


Figure 4-9 Probability plots of the sums of the monthly aggregated time series of the spare parts demand for the eight pieces of equipment

Table 4-5 presents average statistics of the 300 monthly aggregated item level spare parts demand time series. Although, the data ranging between January 2002 and November 2007 were obtained, the average statistical features of the monthly aggregated data were measured throughout periods between January 2002 and December 2006. This was in order to present the statistics consistently with quarterly and yearly aggregated data as shown in Table 4-6. This is because different data aggregation methods (i.e. monthly, quarterly and yearly aggregations) are difficult to compare with only 11 months of data periods for the year 2007. In order to produce summary statistics, the statistical features of individual item level time series were averaged over each piece of equipment. The correlation of an item level time series was measured with other item level time series in the group which was constructed as described in Subsection 4.3.2.

Table 4-5 Average statistics for the monthly aggregated data

	Gun I	Gun II	Gun III	ME I	ME II	GE I	AC I	RD I	Total
Slope	-0.05	**-0.01**	-0.02	-0.27	**-0.46**	-0.18	-0.04	-0.03	-0.29
Corr(item)	0.44	**0.22**	**0.54**	0.50	0.26	0.44	0.29	0.44	0.36
Corr(group)	0.82	0.77	**0.87**	0.82	**0.73**	0.79	0.77	0.77	0.77
Cv(size)	2.17	2.02	2.30	2.11	2.13	**2.52**	**1.68**	1.89	2.18
Pr(peak)	0.40	0.39	**0.49**	0.36	0.38	0.41	**0.28**	0.33	0.38
Skewness	3.18	**2.85**	3.14	3.87	3.78	**4.77**	3.16	3.63	3.87
Kurtosis	12.93	**11.37**	11.53	19.57	18.77	**28.66**	14.49	19.24	19.78
Pr(zero)	0.61	0.65	**0.72**	0.49	0.45	0.53	**0.38**	0.44	0.49
Mean	2.67	1.02	0.84	12.66	26.30	7.06	4.47	2.39	15.82

Key: the highest scores and the lowest scores are shown in bold.

All of the time series had downward trends as shown by their negative Slopes. Time series for ME II seemed to have the steepest downward trends. In practice there were non-significant trends for ME II indentified as shown in Figure 4-7. The lowest value of Slope for ME II was caused by the three peak demands in 2002 and 2003. There were significant correlations [0.22 ~ 0.54 for Corr(item) and 0.73 ~ 0.87 for Corr(group)].

These identified that the item level time series significantly correlate with each other within the same group and with the group level time series. Time series for Gun III has items which correlate most as shown by the highest Corr(item) and Corr(group). There were high Cv(size) and Pr(peak) in most of the data. This analysis identified that most of the time series were highly variable and highly peaked. Time series for AC I was least variable and least peaked as shown by its lowest Cv(size) and Pr(peak). Time series for GE I was most variable as shown by its highest Cv(size). Time series for Gun III was most peaked as shown by its highest Pr(peak).

There was skewness greater than 2.85 and kurtosis greater than 11.37. This identified that the time series were significantly skewed toward the left and were highly peaked (i.e. leptokurtic distribution) as shown in Table 4-5. Time series for GE I was most skewed as well as most kurtosed; whereas, time series for Gun II was least skewed and least kurtosed as shown in Table 4-5. These features of GE I and Gun II were identified as shown in Figure 4-8. Most of the time series were highly intermittent as shown by a total average Pr(zero) of 0.49. Time series for Gun III was most intermittent, and time series for AC I was least intermittent.

There were features characterised in the equipment groups. Gun III was characterises as the highest Pr(peak), Corr(item), Corr(group) and Pr(zero). GE I was characterised as the highest Cv(size), skewness and kurtosis. AC I was characterised as the lowest Cv(size), Pr(peak) and Pr(zero).

Three mean demand quantity groups were identified as shown in Table 4-5. The small mean demand quantity group (0.84 ~ 2.67) is composed of Gun I, Gun II, Gun III, and

RD. The medium mean demand quantity group (4.47 ~ 7.06) is composed of GE I and AC I. The large mean demand quantity group (12.66 ~ 26.30) is composed of ME I and ME II.

As stated in Subsection 1.2.2, non-normal demand is difficult to forecast (Willemain et al., 1994, Regattieri et al., 2005, Syntetos and Boylan, 2005). The spare parts demand time series were identified as non-normal as shown in Table 4-5. Considering that the time series were non-normal, it was difficult to find a significantly superior forecasting method for an equipment group consisting of a small number of items. A larger sample size produces better estimates of the population (Howitt and Cramer, 2008, Field, 2009). As shown in Table 4-4, some equipment groups (i.e. Gun III and RD I) consist of only six items. In fact, it will be shown in the next chapter that a significantly superior forecasting method could not be found for RD, due to the small group size (i.e. only 6 items). For the purpose of comparing forecasting performance at different equipment groups in the next chapter, the 8 equipment groups were combined into 3 homogeneous equipment groups.

By the functions and the links of the pieces of equipment, the 8 equipment groups were combined into 3 equipment groups (i.e. Gun/RD, ME and GE/AC): a) Guns and RD are composed of complex electrical parts, moved by electric power, and controlled by electric signal. Moreover, some Guns are linked with RD in combat data systems (Saunders, 2009). Therefore, Gun I, Gun II, Gun III and RD were combined into an equipment group, Gun/RD. b) The function of GE I and AC I are mainly assisting MEs. Therefore, GE I and AC I were combined into an equipment group, GE/AC. c) ME I and ME II were combined into an equipment group, ME. This is because both ME I and

ME II are different sizes of a main engine, which substitutes spare parts, and are manufactured by the same manufacturer.

The combining criteria were consistent with the previously mentioned three mean demand quantity groups. Small mean demand quantities were observed in Gun/RD; medium mean demand quantities were observed in GE/AC; and large mean demand quantities were observed in ME. As such, the 3 equipment groups (i.e. Gun/RD, ME, and GE/AC) were established from the 8 pieces of equipment. These 3 equipment groups will be used for comparing the performance of forecasting methods at different equipment groups in the next chapter.

As stated in Subsection 2.4.3, some authors (Kahn, 1998, Dekker et al., 2004) have found that combinatorial forecasting combined with a model that considers seasonality outperforms top-down and direct forecasting. The South Korean Navy generates forecasts based on yearly aggregated data sets (Seon and U, 2009). Yearly aggregated data cannot reflect seasonality. In this research the time series for the 300 items and their 150 groups were aggregated into yearly, quarterly, or monthly aggregations to compare the performance of forecasts produced using these different aggregation approaches.

Table 4-6 presents average statistical features of the demand time series of the 300 items as well as the 150 pairs. The average statistical features of the data were measured throughout periods between January 2002 and December 2006 in order to compare the different data aggregations. In order to produce summary statistics, the statistical features of individual time series were averaged over either each equipment group respectively or the total items.

Table 4-6 Statistical features of the time series

		300 item time series				150 group time series			
		Gun/RD	ME	GE/AC	Total	Gun/RD	ME	GE/AC	Total
Yearly	Corr(item)	0.46	0.37	0.56	0.43	-	-	-	-
	Corr(group)	0.82	0.75	0.82	0.78	-	-	-	-
	Slope	-4.01	-57.02	-19.50	-40.74	-8.01	-114.03	-39.01	-81.48
	Cv(size)	0.68	0.70	0.93	0.75	0.56	0.63	0.90	0.68
	Pr(peak)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Skewness	0.86	0.91	1.42	1.02	0.70	0.98	1.63	1.08
	Kurtosis	0.66	1.00	2.37	1.26	0.21	1.11	2.85	1.37
	Pr(zero)	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Mean	24.11	268.62	79.24	189.83	48.22	537.24	158.49	379.67
Quarterly	Corr(item)	0.43	0.36	0.47	0.39	-	-	-	-
	Corr(group)	0.82	0.77	0.81	0.78	-	-	-	-
	Slope	-0.31	-3.84	-1.41	-2.77	-0.62	-7.66	-2.83	-5.53
	Cv(size)	1.25	1.28	1.44	1.31	1.01	1.13	1.32	1.15
	Pr(peak)	0.29	0.28	0.33	0.20	0.24	0.26	0.30	0.06
	Skewness	1.71	1.94	2.42	2.02	1.43	1.98	2.64	2.05
	Kurtosis	3.58	4.54	7.40	5.05	2.35	4.80	8.72	5.33
	Pr(zero)	0.28	0.17	0.20	0.20	0.11	0.05	0.07	0.06
	Mean	6.03	67.16	19.81	47.46	12.05	134.31	39.62	94.92
Monthly	Corr(item)	0.41	0.33	0.41	0.36	-	-	-	-
	Corr group)	0.81	0.76	0.79	0.77	-	-	-	-
	Slope	-0.03	-0.41	-0.15	-0.29	-0.07	-0.81	-0.30	-0.59
	Cv(size)	2.12	2.13	2.37	2.18	1.69	1.86	2.14	1.90
	Pr(peak)	0.40	0.38	0.39	0.38	0.35	0.33	0.32	0.33
	Skewness	3.16	3.81	4.48	3.87	2.76	3.88	4.89	3.94
	Kurtosis	13.24	19.00	26.16	19.78	10.13	19.99	30.28	20.88
	Pr(zero)	0.61	0.46	0.50	0.49	0.44	0.26	0.30	0.29
	Mean	2.01	22.39	6.60	15.82	4.02	44.77	13.21	31.64

For all the equipment groups, Corr(item) slightly increased in order of monthly, quarterly and yearly. However, Corr(group) were similar regardless of the method of data aggregation. Slope, Cv(size), Pr(peak), skewness, kurtosis, and Pr(zero) for all the equipment groups declined in order of monthly, quarterly and yearly. While skewness and kurtosis were pronounced in quarterly and monthly time series, yearly time series for all the equipment groups were less skewed and less kurtosed. Especially, Pr(peak) and Pr(zero) in yearly aggregated time series for all the equipment groups were zero with the exception of Pr(zero) 0.01 for the item level time series of Gun/RD. This seemed to indicate that the yearly aggregated time series do not represent non-normal demand features anymore. However, as yearly time series have very few data points (i.e. five observations), it is problematic to assert that yearly time series are either normal data or closer to normal data.

When comparing the features of group level time series with the features of item level time series, some patterns were identified. Compared to item level time series, skewness and kurtosis increased in group level time series with the exception of Gun/RD; whereas Slope, Cv(size), Pr(peak) and Pr(zero) decreased in group level time series. Data aggregation in group level time series might simply lead to strengthening skewness and kurtosis in ME and GE/AC and Slope in all the equipment groups.

The decreased 3 statistical features [i.e. Cv(size), Pr(peak) and Pr(zero)] in group level time series should be noted. As stated, Cv(size) and Pr(peak) might represent the degree of variability capturing the erratic or lumpy demand, and Pr(zero) might represent the degree of intermittency capturing the intermittent, slow moving, clumped or lumpy demand. As such, reduced non-normal demand features at group level time series were characterised. This reduced non-normality of group level time series suggests that hierarchical forecasting would be superior to direct forecasting (Gross and Sohl, 1990, Fliedner and Lawrence, 1995, Fliedner, 2001).

Some relative demand features in the equipment groups were identified. Gun/RD was characterised as having higher intermittency and smaller demand volume, owing to higher Pr(zero) and smaller Mean. ME was characterised as having lower correlation, steeper downward trend, lower intermittency, and larger demand volume owing to lower

Corr(item), lower Corr(group), steeper Slope, lower Pr(zero), and greater mean. GE/AC was characterised as having higher variability, greater peakedness, and greater deviation from a normal distribution owing to higher Cv(size), higher Pr(peak), higher skewness, and higher kurtosis.

4.3.5 Seasonality

As stated in Subsection 2.5.2, seasonality could influence the performance of hierarchical forecasting when seasonality exists in the time series (Kahn, 1998, Dekker et al., 2004). As suggested by Figure 4-4, spare parts demand in the South Korean Navy could be influenced by seasonal effects. For example, during winter seasons, more Naval warships are scheduled to be overhauled than the other seasons because of the reduced demand of the Naval warships caused by bad sea conditions. On the other hand, during the fishing season in summer and also in autumn when it is more appropriate for exercises due to better sea conditions, operational demands on Naval warships are likely to be high. Thus, major overhauls are unlikely to be scheduled.

An approximate value for the standard error of seasonal effect, $S.E.(\hat{s}_j)$, was employed to measure the significance of seasonal effects as shown in equation (4-5). In this research, the value, $\ddot{y}_{ij}$, was adjusted by linear trend and additive seasonality. The process of trend and seasonal adjustment will be described later in this chapter.

$$S.E.(\hat{s}_j) \cong stdev(\ddot{y}_{ij}) / \sqrt{n_j} \qquad (4\text{-}5)$$

where:

$\hat{s}_j$ = the seasonal effect

$\ddot{y}_{ij}$ = the value adjusted for trend and seasonality

n_j = the number of observations of the j^{th} seasonal deviation

A t-test compares H_0 and H_1 (H_0: $\hat{s}_j = 0$ versus H_1: $\hat{s}_j \neq 0$). t values were calculated for each value of j as shown in equation (4-6) and compared each with t_{n_j-1}.

$$|t| = \frac{|\hat{s}_j|}{S.E.(\hat{s}_j)} \tag{4-6}$$

Quarterly seasonality was tested throughout periods between the 1st quarter of 2002 and the 3rd quarter of 2007. Table 4-7 presents the summary of the test for the significance of seasonal effect (Seffect) in the sum of the 300 quarterly aggregated time series for the 300 items. The seasonal effects were calculated based on the data adjusted by linear trend. No seasonal effect was significant.

Table 4-7 Test for quarterly seasonal effect

	1st	2nd	3rd	4th
Seffect	4039.8343	694.5145	-4389.3053	-414.0522
Std	11784.60061	7561.736001	5031.058251	2489.975468
n_j	6	6	6	5
S.E.	4811.043052	3087.065795	2053.92093	1113.550882
$\|t\|$	0.839700308	0.224975607	2.137037151	0.371830517
n_j-1	5	5	5	4
p-value	0.439	0.831	0.086	0.729

Table 4-8 presents quarterly seasonal effects in each equipment group with their p-values. All quarterly seasonal effects in all equipment groups were non-significant.

Table 4-8 Quarterly seasonal effect in equipment groups

		1st	2nd	3rd	4th
Gun/RD	Seffect	85.4	27.8	-34.3	-94.6
	p-value	0.398	0.625	0.571	0.088
ME	Seffect	3144.1	846.5	-3938.7	-62.2
	p-value	0.436	0.786	0.082	0.964
GE/AC	Seffect	810.3	-179.7	-416.3	-257.2
	p-value	0.549	0.663	0.153	0.523
Total	Seffect	4039.8	694.5	-4389.3	-414.1
	p-value	0.439	0.831	0.086	0.729

Monthly seasonality was tested throughout periods between the January 2002 and November 2007. Table 4-9 presents the summary of the test for the seasonal effects of the sum of the 300 monthly aggregated data for the 300 items. The seasonal effect of July was the only significant seasonal effect as p–value < 0.05. This might be explained by the higher demand of Naval warships which reduces major overhauls in July.

Table 4-9 Test for monthly seasonal effect

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Seffect	-1501.0	49.3	5527.5	-640.3	-1191.6	2539.2	**-1563.6**	-1399.9	-1436.2	-1652.5	-245.0	**1816.8**
Std	1946.1	4309.9	11288.3	2828.2	1443.4	7626.4	1402.4	1963.5	2335.6	1792.1	1798.5	3456.7
n_j	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	5.0
S.E.	794.5	1759.5	4608.4	1154.6	589.2	3113.5	572.5	801.6	953.5	731.6	734.2	1545.9
$\|t\|$	1.9	0.0	1.2	0.6	2.0	0.8	2.7	1.7	1.5	2.3	0.3	1.2
n_j-1	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	4.0
p-value	0.118	0.979	0.284	0.603	0.099	0.452	0.041	0.141	0.192	0.074	0.752	0.305

The significant seasonal effect is shown in bold.

Table 4-10 presents monthly seasonal effects in each equipment group with their p-values. Some monthly seasonal effects were significant. The highly negative seasonal effects in July for the sum of the 300 monthly time series were also pronounced in Gun/RD and ME.

Table 4-10 Monthly seasonal effect in equipment groups

		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Gun/RD	Seffect	-31.9	11.5	105.3	10.7	15.6	0.4	**-44.5**	-9.6	18.1	**-39.0**	-0.2	**-43.7**
	p-value	0.066	0.648	0.360	0.777	0.631	0.993	0.033	0.693	0.522	0.025	0.993	0.007
ME	Seffect	-1292.2	120.5	4355.7	-684.9	-1072.9	2626.1	**-1327.0**	-1247.8	-1360.1	-1590.1	-155.2	1953.4
	p-value	0.144	0.943	0.244	0.484	0.076	0.430	0.042	0.142	0.170	0.060	0.848	0.276
GE/AC	Seffect	-176.9	-82.7	1066.5	33.9	-134.3	-87.3	-192.0	-142.5	-94.2	-23.5	-89.6	-92.9
	p-value	0.139	0.633	0.426	0.900	0.245	0.476	0.068	0.214	0.486	0.832	0.421	0.598
Total	Seffect	-1501.0	49.3	5527.5	-640.3	-1191.6	2539.2	**-1563.6**	-1399.9	-1436.2	-1652.5	-245.0	1816.8
	p-value	0.118	0.979	0.284	0.603	0.099	0.452	0.041	0.141	0.192	0.074	0.752	0.305

Significant seasonal effects are shown in bold.

For Gun/RD, two more seasonal effects were significant. There was a significantly negative seasonal effect of Gun/RD in October. This might be explained by the high operational demands on Naval warships caused by better sea conditions. The significantly negative seasonal effect of Gun/RD in December is difficult to explain. A positive seasonal effect in December was expected because of increased major overhauls during the winter season. There was a positive seasonal effect for ME in December, although the seasonal effect was non-significant. This inconsistency could be explained by demand information distortion caused by many sources. The demand information distortion will be discussed in more detail later in this chapter.

4.3.6 Procurement lead time

As stated in Subsection 2.5.2, Shlifer and Wolff (1979) suggested that a long forecasting horizon makes the performance of top-down forecasting better than direct forecasting. This required the examination of the procurement lead time (PROLT) of the Naval spare parts. Figure 4-10 and Table 4-11 present the features of the PROLT. Although the PROLTs ranged from 3 months to 18 months, 70.3% of the PROLTs (211 items) were concentrated on the mode (i.e. 10 months) for all 300 spare parts as shown in Figure 4-10. High concentration on the mode was identified by the high value of kurtosis (i.e.

8.64 in total). 16.7% of items (50 items) had 9 months PROLT, and 2.7% of items (8 items respectively) had 8, 6 and 5 months PROLT. The PROLTs were concentrated toward to the left from the mode value with the exception of Gun/RD (as shown in Figure 4-10 and Table 4-11). Gun/RD had shorter PROLT (mean 9.2 months) than ME and GE/AC and spread widely (standard deviation is 2.67 months) as shown in Figure 4-10.

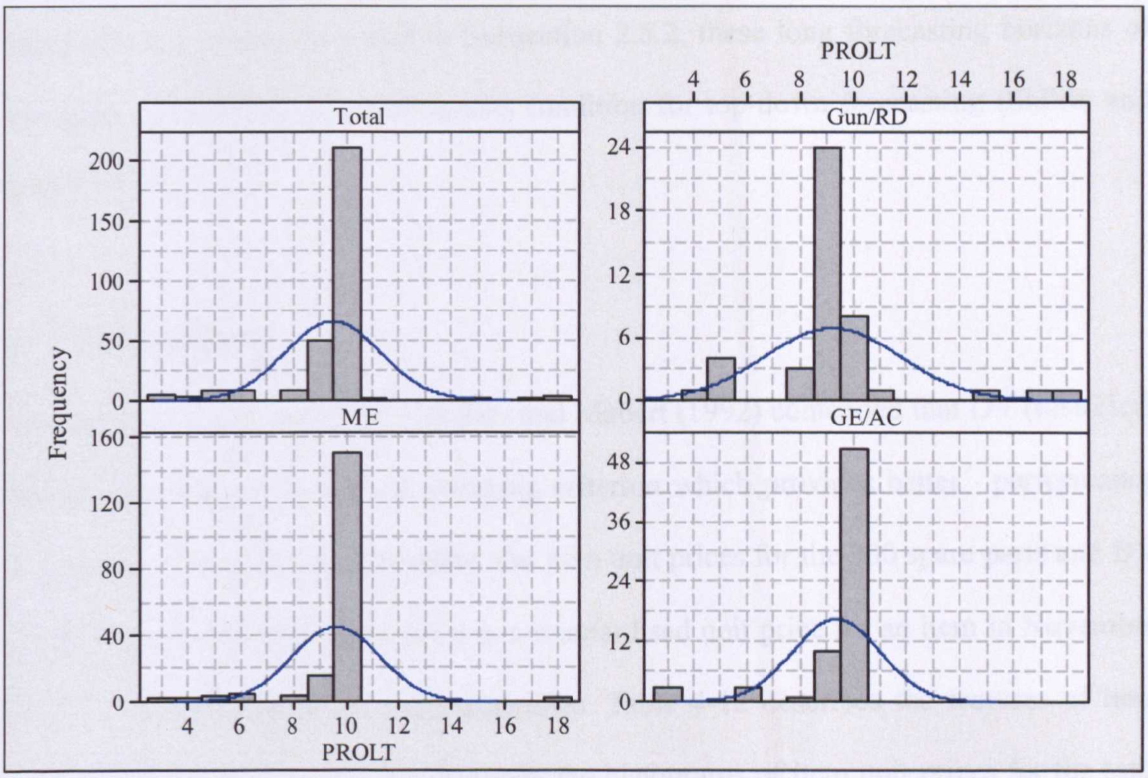


Figure 4-10 Histograms of PROLT

Table 4-11 PROLT analysis

	Mean	Median	Std	Skewness	Kurtosis
Total	9.47	10	1.81	-0.33	8.64
Gun/RD	9.20	9	2.57	1.32	4.68
ME	9.57	10	1.65	-0.61	11.18
GE/AC	9.34	10	1.64	-3.04	8.79

Longer PROLT increases RO1 (Requirement Objective), because it covers longer periods of demand. A large volume of RO1 (composed of the PROLT, the PC and the

SL) introduces large stock holdings. Large stock holdings amplify the disadvantages of stock holdings such as obsolescence, damage, deterioration and loss (Slack et al., 2004). This increases the inventory carrying costs such as investment in the inventory, warehousing and holding costs (Waller, 2003).

The long PROLTs indicated that the forecasting horizons need to be quite long periods ranging from 15 months to 30 months including 12 months review periods (procurement cycle). As stated in Subsection 2.5.2, these long forecasting horizons of the spare parts imply an advantageous condition for top-down forecasting (Shlifer and Wolff, 1979).

4.3.7 Item unit price

As stated in Subsection 2.5.2, Fliedner and Mabert (1992) contended that DV (historical dollar volume) is a significant grouping criterion which provides better performance for top-down forecasting. Therefore, the item unit prices for the 300 spare parts and DV were examined. An item unit price is a standardised unit price for an item in November 2007. It ignores minimum batch quantities. Table 4-12 describes the features of item unit prices and DV. Figure 4-11 presents the histograms of item unit prices for the total items and the items in each equipment groups. Figure 4-12 presents the histograms of DVs for the total items and the items in each equipment groups. DV was calculated as "DV= the sum of historical demand for an item between January 2002 and November 2007 × item unit price" for Table 4-12 and Figure 4-12.

Table 4-12 Item unit price and DV (Unit: ₩1,000)

		Mean	Median	Std	Skewness	Kurtosis
Unit Price	Gun/RD	781	637	925	3.40	15.78
	ME	171	8	1,004	10.31	114.14
	GE/AC	129	4	391	4.62	22.91
	Total	251	11	915	9.13	103.30
DV	Gun/RD	69,348	40,950	106,566	4.48	23.63
	ME	72,068	3,705	263,114	5.87	37.74
	GE/AC	19,932	868	45,264	3.74	15.26
	Total	59,852	4,913	214,147	7.03	56.23

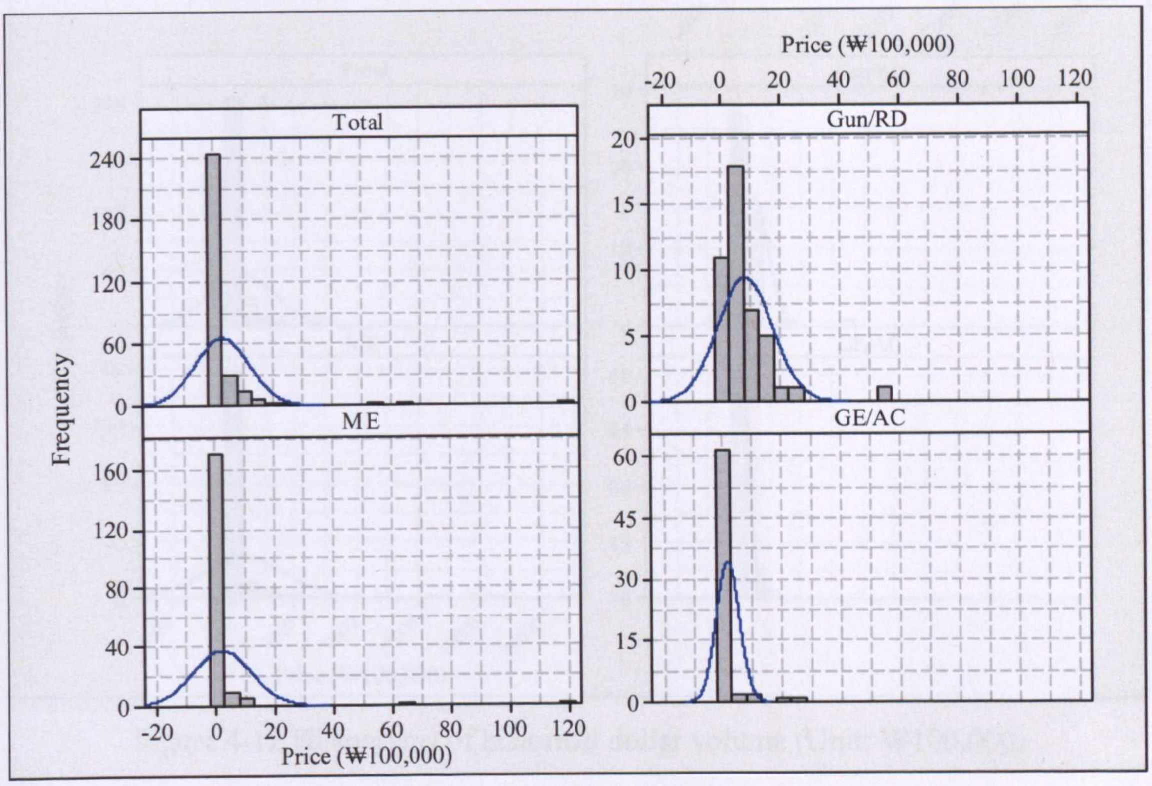


Figure 4-11 Histograms of item unit price (Unit: ₩100,000)

Most of the unit prices for the items were smaller than ₩500,000 (£255) and the mean unit price was ₩251,000 (£128). Spare parts for Gun/RD were characterised as more expensive spare parts as the mean unit price of the spare parts for Gun/RD was more expensive [i.e. ₩781,000 (£399)]. The higher prices of the spare parts for Gun/RD than other equipment groups were identified in Figure 4-11. Most of the DVs were smaller than ₩150,000,000 (£76,648) and the mean DV was ₩59,852,000 (£30,583). Owing to the larger historical demand volume (i.e. Mean) of the spare parts for ME as shown in

Table 4-6, the mean DV of ME was greater than the mean DV of Gun/RD. However, the mean DV of ME was not a good statistic to represent the DV of ME, as the DVs of ME spread widely [standard deviation of ME is ₩263,114,000 (£134,448)]. The wide-spreading DVs of ME caused by outliers are identified in Figure 4-12.

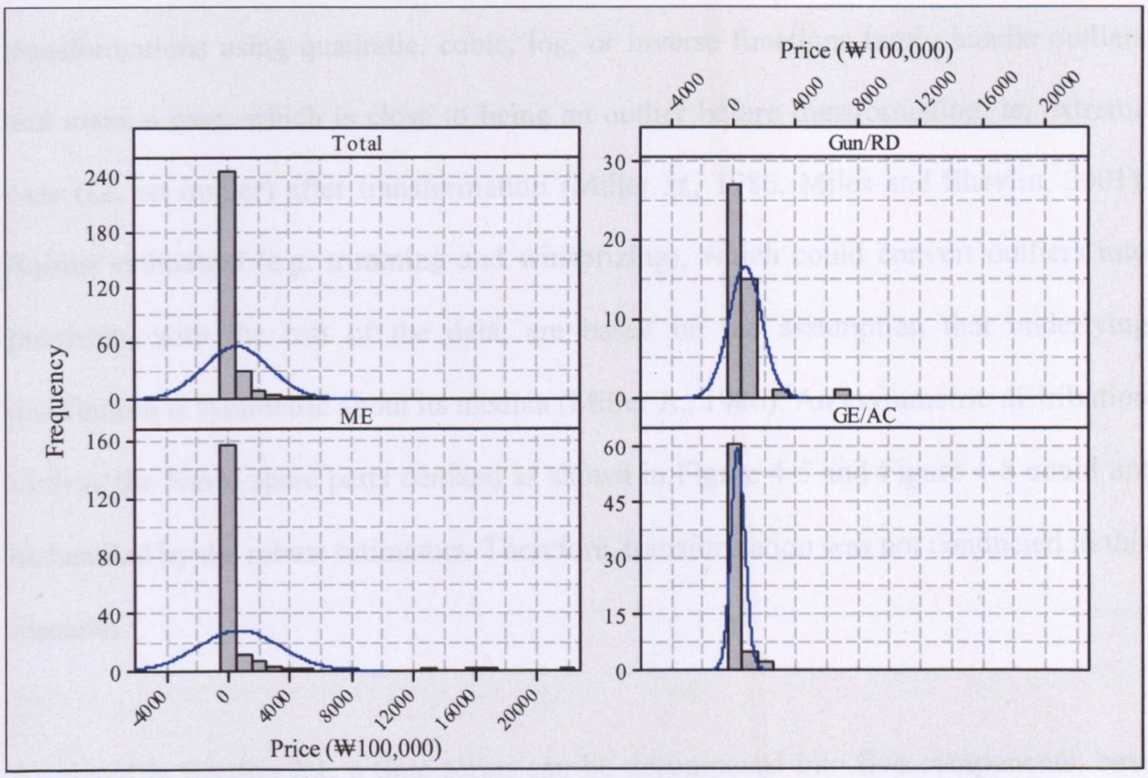


Figure 4-12 Histograms of historical dollar volume (Unit: ₩100,000)

In this section, the demand features of the 300 spare parts data in the South Korean Navy were identified using various measures. The time series of the spare parts demand were found to be non-normal. Downward trends and non-significant seasonality in most of the time series were identified. Some advantageous conditions for hierarchical forecasting were also identified such as reduced non-normal demand features at group level time series than item level time series and long procure lead times.

4.4 Decomposition Procedure

The time series of the Naval spare parts demand were found to be non-normal. The problems of transformation were discussed in Subsection 2.2.1. Transformation has limitations (Chatfield, 2004) and the forecasting performance improvement for non-normal demand was found to be very small (Nelson and Granger, 1979). The linear transformations using quadratic, cubic, log, or inverse functions rarely handle outliers and make a case, which is close to being an outlier before transformation, an extreme case (i.e. an outlier) after transformation (Miller Jr., 1986, Miles and Shevlin, 2001). Robust estimators (e.g. trimming and winsorizing), which could convert outliers into proximity with the rest of the data, are based on the assumption that underlying distribution is symmetric about its median (Miller Jr., 1986). An asymmetric distribution such as the Naval spare parts demand as shown in Figure 4-5 and Figure 4-8 could not be handled by the robust estimators. Therefore, transformation was not conducted in this research.

As stated in Section 2.1, a time series can be decomposed into five components: base demand, seasonal, trend, cyclic, and irregular components. In order to examine more closely other components, trend and seasonal components can be estimated or removed (Hyndman et al., 1998, Chatfield, 2004). Simple exponential smoothing will be employed to generate forecasts for both hierarchical and direct forecasting methods in the next chapter. In order to employ simple exponential smoothing, as stated in Subsection 2.2.3, trend and seasonal components have to be measured or removed (Waters, 1991, Gardner Jr. and Diaz-Saiz, 2002). Thus, this research employed a decomposition process. The decomposition process conducted for the trend and seasonal components in the spare parts demand time series are described as follows.

Some authors examined various trends such as polynomial trends (Morris and Glassey, 1962) and linear/exponential trends (Gardner, 1985) for exponential smoothing. In their analytical study, Morris and Glassey (1962) contended that exponential smoothing models coupled with polynomial trends have a tendency to amplify noise in the time series. In his analytical study, Gardner (1985) argued that exponential smoothing models coupled with exponential trends are inaccurate at long forecasting horizons, however, a linear trend is typically used for any forecasting horizon. Exponential trends are not suitable for the spare parts demand time series in the South Korean Navy, because the Navy requires a long forecasting horizon as stated in Subsection 4.3.6.

A linear trend is defined as in equation (4-7) (Cryer and Chan, 2008, p. 27). This research employed the linear trend in order to extrapolate this trend to produce forecasts. β_0 and β_1 in equation (4-7) were estimated by a regression method that minimises equation (4-8) (Cryer and Chan, 2008). Peak points in 2002 ~ 2003 and downward trends were identified in the spare parts demand time series as shown in Figure 4-4 and Figure 4-7. A time series can be expressed as in equation (4-9). The linear trend, $\breve{y}_t$, was removed from the data in this research so as to analyse local fluctuations.

$$\breve{y}_t = \beta_0 + \beta_1 t \tag{4-7}$$

where: $\breve{y}_t$ = the trend value at time t

$$Q(\beta_0, \beta_1) = \sum_{t=1}^{n} [y_t - (\beta_0 + \beta_1 t)]^2 \tag{4-8}$$

where: y_t = the observation at time t

n = the number of time periods

$$y_t = \breve{y}_t + \varepsilon_t \tag{4-9}$$

where: y_t = the observation at time t

$\breve{y}_t$ = the trend value at time t

ε_t = the random error at time t

As stated in Subsection 2.3.2, when the size of seasonality is directly proportional to the mean, the seasonality is referred to as multiplicative (Chatfield, 2004). The seasonality of the Naval spare parts is unlikely to be directly proportional to the mean as shown in Figure 4-4 and Figure 4-7. As stated in Subsection 2.2.3, Gardner Jr. and Diaz-Saiz (2002) found that the forecasting performance of exponential smoothing models coupled with additive decomposition was significantly superior to the performance of exponential smoothing models coupled with mulplicative decomposition for predicting automotive spare parts demand that was intermittent. The Naval spare parts demand time series were found to be highly intermittent as the proportion of zero demand periods [Pr(zero)] was high as shown in Table 4-6. Therefore, this research employed an additive seasonality model.

The seasonal deviations were obtained as historical demand values deducted by trend values:

$$\dot{y}_t = y_t - \breve{y}_t \tag{4-10}$$

where: $\dot{y}_t$ = the seasonal deviations at time t

$\breve{y}_t$ = the trend value at time t

The seasonal effects, $\hat{s}_j$, were calculated as the seasonal means deducted by overall means:

$$\hat{s}_j = \bar{y}_{.j} - \bar{y}_{..} \tag{4-11}$$

where: $\hat{s}_j$ = the seasonal effect in the j^{th} season

$\bar{y}_{.j}$ = the seasonal means in the j^{th} season

$\bar{y}_{..}$ = the overall means

Then, the seasonally adjusted values were calibrated as historical demand values deducted by seasonal effects:

$$\ddot{y}_{ij} = y_{ij} - \hat{s}_j \tag{4-12}$$

where: $\ddot{y}_{ij}$ = the seasonally adjusted value in the j^{th} season in the i^{th} year

y_{ij} = the observation in the j^{th} season in the i^{th} year

Trend and seasonally adjusted values were produced as seasonally adjusted values deducted by trend values:

$$\dddot{y}_{ij} = y_{ij} - \hat{s}_j - \breve{y}_{ij} \tag{4-13}$$

where: $\dddot{y}_{ij}$ = the trend and seasonally adjusted value in the j^{th} season in the i^{th} year

$\hat{s}_j$ = the seasonal effect in the j^{th} season

$\breve{y}_{ij}$ = the trend value in the j^{th} season in the i^{th} year

Each time series for the demand for the 300 items (aggregated by the yearly, quarterly or monthly aggregation method) was adjusted by the above decomposition procedure. The adjusted time series were aggregated across the 300 items used for analysis. Figure 4-13, Figure 4-14 and Figure 4-15 describe the aggregated time series plots across the adjusted or unadjusted yearly, quarterly or monthly aggregated time series respectively. The downward trends were removed in the trend adjusted data plots and the trend and seasonality adjusted data plots. The removal of the downward trends identified that the decomposition characterised the trends. However, the decomposition failed to remove seasonality, as most of the seasonal effects were non-significant as shown in Subsection 4.3.5. The two peak points were observed in all the aggregated time series plots. The smallest positive peaks were observed in the trend and seasonal adjusted quarterly data plot.

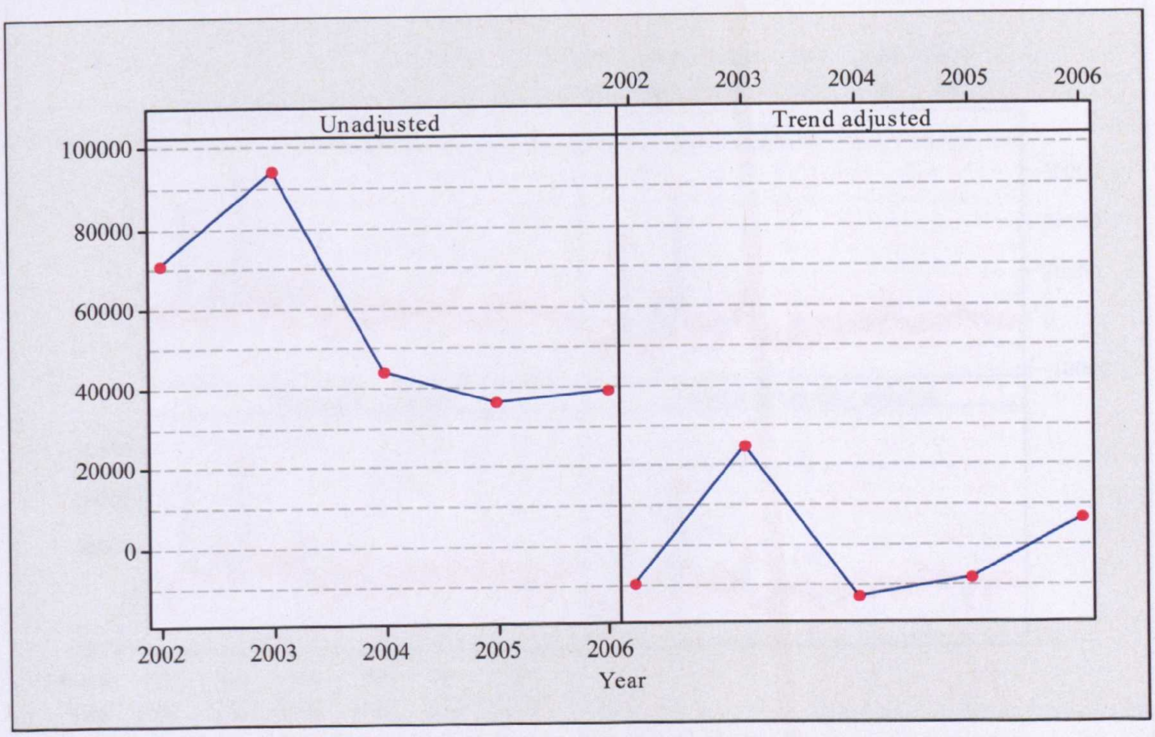


Figure 4-13 Yearly data adjustment

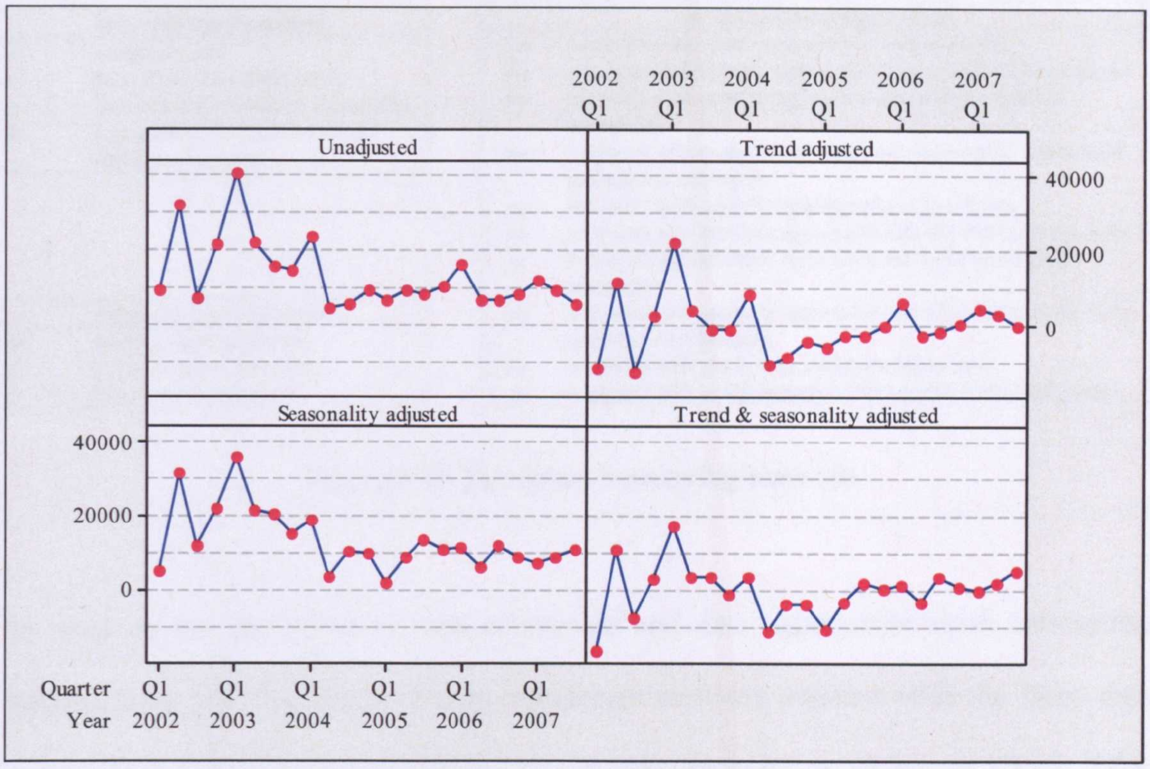


Figure 4-14 Quarterly data adjustment

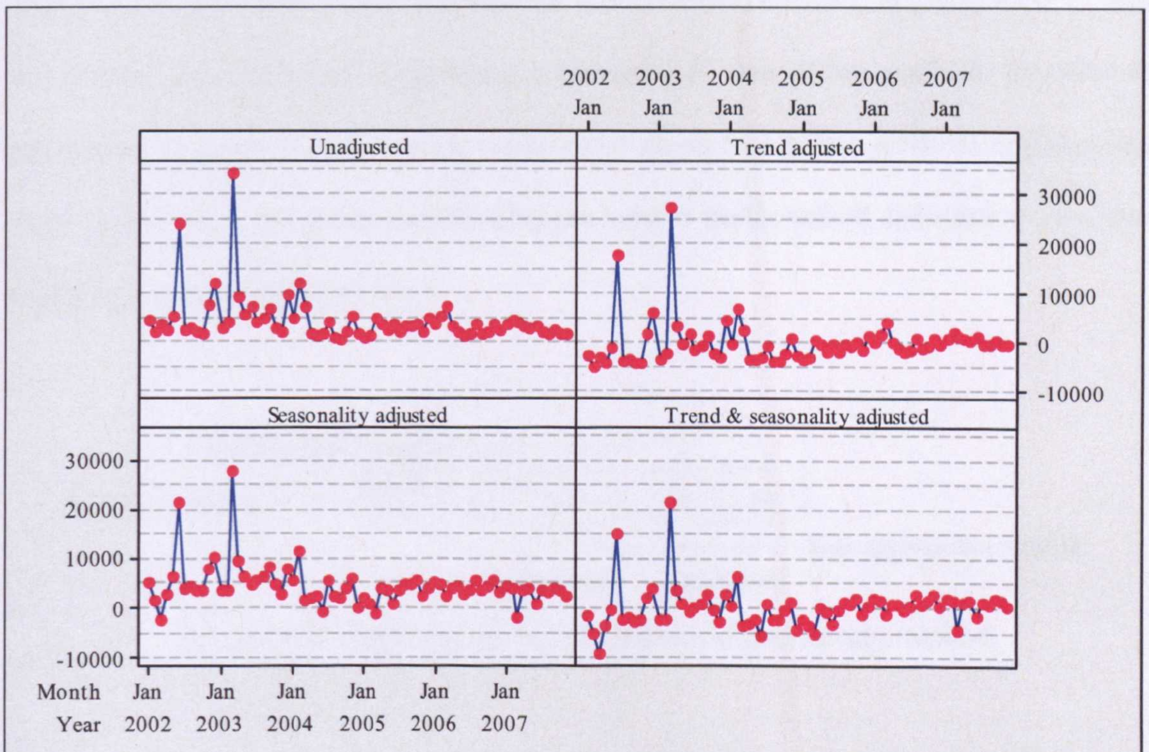


Figure 4-15 Monthly data adjustment

Data adjustment method	
u	unadjusted data
t	data adjusted for linear trend
s	data adjusted for additive seasonality
ts	data adjusted for linear trend and additive seasonality

Data aggregation method	
m	monthly aggregated data
q	quarterly aggregated data
y	yearly aggregated data

10 direct forecasting methods	
um	a forecast with monthly aggregated unadjusted data
tm	a forecast with monthly aggregated data adjusted for linear trend
sm	a forecast with monthly aggregated data adjusted additive seasonality
tsm	a forecast with monthly aggregated data adjusted for linear trend and additive seasonality
uq	a forecast with quarterly aggregated unadjusted data
tq	a forecast with quarterly aggregated data adjusted for linear trend
sq	a forecast with quarterly aggregated data adjusted additive seasonality
tsq	a forecast with quarterly aggregated data adjusted for linear trend and additive seasonality
uy	a forecast with yearly aggregated unadjusted data
ty	a forecast with yearly aggregated data adjusted for linear trend

Figure 4-16 Ten direct forecasting methods

In order to test the effect of data adjustment and data aggregation upon forecasting performance, the four kinds of data adjustment methods together with the three data aggregation methods were employed to generate forecasts as shown in Figure 4-16. Abbreviations were utilised to refer to these data succinctly. Ten direct forecasting methods are generated. In the next chapter, forecasts are generated at group level as well as at item level in order to produce a hierarchical forecasting method. In order to represent a hierarchical forecasting method two direct forecasting methods at group and item levels are to be expressed in sequence with a prefix which indicates a proration method as shown in Figure 4-17.

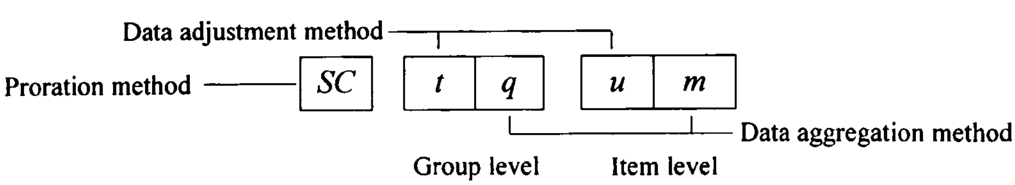


Figure 4-17 Abbreviation for a hierarchical forecasting method

The exemplified forecasting method, *SCtqum* indicates a forecasting method generated by a simple combination (SC) between the forecast with quarterly aggregated data adjusted for linear trend at group level (*tq*) and the forecast with monthly aggregated unadjusted data at item level (*um*). Proration methods employed for this research are described in the next chapter in detail.

4.5 Sources of Non-Normality

Under a holistic single case design (Yin, 2003), a typical military logistical case (i.e. the case of spare parts in the South Korean Navy) was employed to verify the theory which was developed in this research. As mentioned in Subsection 3.6.1, construct validity tests correct operational measures for the concepts being studied (McCutcheon and Meredith, 1993, Yin, 2003). The historical spare parts demand data obtained from the logistical database of the Naval Logistics Command (NLC) were thought of as establishing the construct validity of this research. However, the non-normal demand causes distrust in the data generating process. In practice, the low reliability of data was contended to be the major reason of the low accuracy of forecasting spare parts demand in the South Korean Navy (Seon and U, 2009). Several sources of the non-normal demand can be presented as follows.

4.5.1 A few large customers

A few large customers could induce non-normality (Silver, 1970, Eaves, 2002). Irregular large orders from a few customers (e.g. Naval warships) could be highly sporadic. For instance, only 9 Ulsan class Frigates and 28 Po-Hang class Corvettes have been commissioned in the South Korean Navy (Saunders, 2009). Spare parts demand for the 76 MM guns which are uniquely installed in these warships has a tendency to fluctuate depending on irregular large demand volumes from some of these warships.

4.5.2 Operating factors

Ghobbar and Friend (2002, 2003) pointed out three operating factors (i.e. aircraft utilisation rate; component overhaul life; and primary maintenance process) which contributed to generating non-normality of data in their experimental study with aircraft spare parts data. Aircraft spare parts demand has similar characteristics to Naval spare parts demand in that it has non-normal demand patterns and has to respond to the peak demand reasonably (Ghobbar and Friend, 2002).

Ghobbar and Friend (2002, 2003) described the three factors as follows. Aircraft utilisation rate was expressed as 'hours (cycles) per period' form (e.g. 7.1 hours per day or 3.5 cycles per day). Over-utilisation can cause costly mechanical failure as well as shorter asset-life; under-utilisation can lead to a reduction in fleet size. Component overhaul life was expressed as 'flying hours/landings' controlled by mean time between overhaul (MTBO) so as to assist the movement of spare parts into periodic maintenance tasks smoothly and systematically. Primary maintenance process is a categorical factor which is divided into three sub-factors such as hard-time, on-condition and condition-monitoring. Hard-time indicates a preventive process in which the known deterioration of an item is maintained to an acceptable level by the maintenance actions carried out at periods related to time in service (i.e. calendar time, number of cycles, or number of landings); on-condition indicates a preventive process which requires periodical inspection against some appropriate physical standard; and condition-monitoring indicates a corrective procedure based on information on items (e.g. the thickness of a brake pad).

Ghobbar and Friend (2002) noted that all the three factors affected demand variation (i.e. the square coefficient of variation in spare parts demand sizes); and aircraft utilisation rate was a major source of non-normality in the way that, as it increases, the square coefficient of variation in demand size increases and the average inter-demand interval decreases respectively. These operating factors can also be one of the main sources for non-normality of the spare parts demand in the South Korean Navy.

By way of example, as stated in Subsection 4.3.4, the large demand for spare parts in June of 2002 and March of 2003 in Figure 4-4 could be explained by an increased demand of the warships caused by the sea battles with the North Korean Navy in June of 2002 (Jie-Ae, 2002) and subsequent preparation against possible clashes before the fishing season in the next year 2003. As stated in Subsection 2.7.3, this kind of unexpected demand is the reason why the Navy holds a large amount of stock.

4.5.3 Multi-echelon inventory systems

The above two sources of the non-normal demand are unavoidable, because these reflect the true demand fluctuations. However, the following three sources reflecting the difficulty of obtaining true demand data might cause biasing effects on forecasts.

'Bullwhip effect' is defined as "the phenomenon where orders to the supplier tend to have larger variance than sales to the buyer (i.e. demand distortion), and the distortion propagates upstream in an amplified form (i.e. variance amplification)" (Lee et al., 1997, p. 546). Lee et al. (1997) claimed that the 'bullwhip effect' can occur in a wholesaler-distributor relationships. When an inventory manager in the supply chain experiences a surge of demand in one period, the manager is likely to interpret it as a signal of high future demand, adjust the demand forecast, and place a large order. Typically, an upstream supplier relies on the order data from the downstream retailer. Lee et al. (1997) contended that this phenomenon will cause the supplier to lose track of the true demand pattern at the retail end. Likewise, in the multi-echelon inventory systems in the South Korean Navy, the demand information is likely to be distorted.

Choi et al. (2005) illustrated information distortion with respect to the inventory control method based on ASL in the multi-echelon systems. If an N-ASL item changes to ASL with a demand quantity (e.g. 100 units) in the year *Y*, a retail unit would place an order to raise inventory position to Requisition Objective (RO2), 127 units, which is more than Operating Level (OL), 100 units. In this case, RO2 consists of OL (100 units) and SL + OST (27 units). Therefore, the demand record of the year *Y* (127) is followed on. If the item is kept as ASL in the year *Y*+1 and later on, the retail unit would place an order to raise the inventory position to RO2 which is the same amount of stock as OL; that is 100 units (see Figure 4-18).

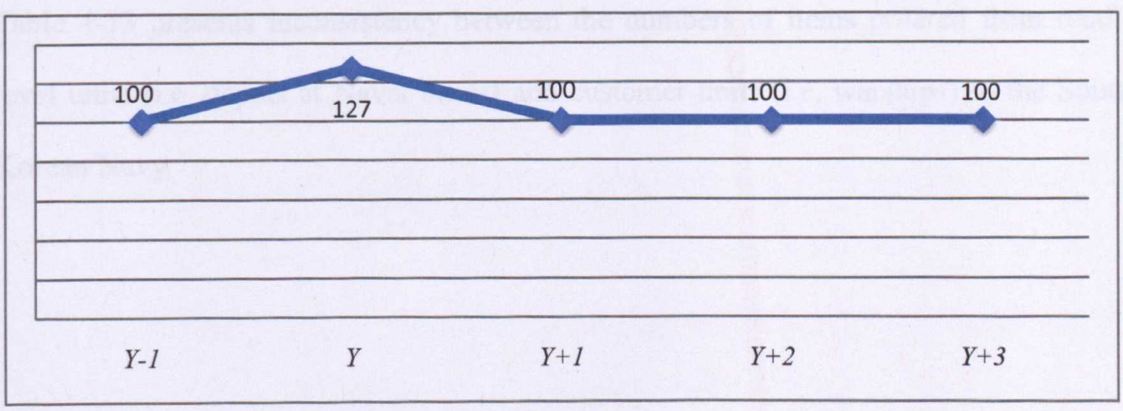


Figure 4-18 Change of demand (N-ASL → ASL) (Choi et al., 2005)

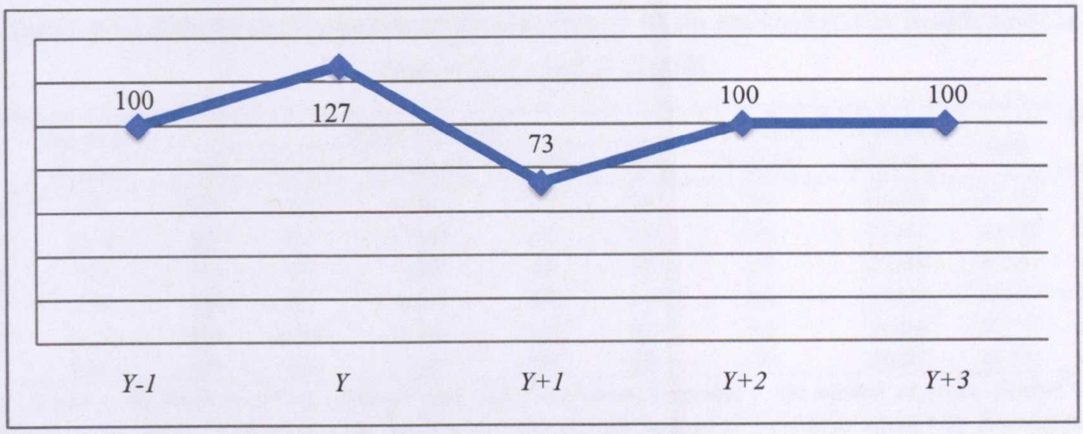


Figure 4-19 Change of demand (N-ASL → ASL → N-ASL) (Choi et al., 2005)

However, if the item changes to N-ASL in the year *Y*+1, the retail unit would place an order only 73 units: i.e. order quantity (73 units) = OL (100 units) – stock on hand (27 units) (see Figure 4-19). This is because an N-ASL item does not need SL and OST. Although customers (i.e. warships) consume the same amount of spare parts (100 units), fluctuation in the demand record could be observed. Choi et al (2005) contended that in order to prevent this unnecessary fluctuation, the records of RO2 and OL have to be managed separately.

Table 4-13 presents inconsistency between the numbers of items ordered from retail-level units (i.e. depots at Naval bases) and customer units (i.e. warships) in the South Korean Navy.

Table 4-13 Inconsistency between the number of items ordered in the South Korean Navy (Seon and U, 2009)

	Customer>0	Retail / Customer					Customer=0	Customer=0	Sum	(b+c)
	Retail=0	<50%	50~80%	80~120%(b)	12~150%	150%<	Retail>0	Retail=0 (c)		/ sum
'02	878	572	1,125	22,821	291	257	320	19,293	45,557	92%
'03	1,311	643	991	17,409	243	246	202	24,512	45,557	92%
'04	785	397	736	17,300	368	323	281	25,367	45,557	94%
'05	1,566	544	946	13,819	405	544	401	27,332	45,557	90%
'06	2,113	979	1,086	12,300	333	419	369	27,958	45,557	88%
'07	2,043	850	916	11,569	387	496	249	29,047	45,557	89%

Key: Retail = the number of items ordered from retail-level units; Customer = the number of items ordered from customer units: Retail / Customer = the number of items ordered from retail-level units divided by the number of items ordered from customer units.

Requisition Objective (RO2) of retail-level units is approximately inventory for two months (Seon and U, 2009). This is roughly 20% of annual demand quantity. Seon and U (2009) assumed that the ratios between 80 ~ 120% are acceptable differences in orders between retail-level units and customer units. When (b) and (c) are correct cases, the ratio of correct cases can be calculated as "(b + c)/sum" and are presented in the last column. Approximately 9% of cases were incorrect based on this criterion. This implies 9% of demand data were distorted.

As stated, an order of an ASL item can come from either the purpose of repair and maintenance or the purpose of keeping inventory position. Whilst the order from the purpose of repair and maintenance is true demand, the order from the purpose of keeping inventory is proxy demand. The problem is that forecasts for procurement decision are generated using historical order data of retail-level units (not customers) which include orders from both purposes (Seon and U, 2009). These could not identify true demand. This is a limitation of this research.

4.5.4 Budgeting process

Choi et al. (2005) claimed that the budgeting process could cause non-normality in the

Military. There is a tendency of inventory managers to under-request spare parts when the budget is deficient; however, over-request spare parts when there is sufficient budget. This may amplify the information distortion with the multi-echelon inventory systems.

4.5.5 Maintenance system

Choi et al. (2005) claimed information distortion caused by the maintenance system could induce non-normality of the demand. Substituting spare parts during maintenance processes could be one of the reasons for non-normality. As mentioned in Subsection 2.5.2, when two items are highly substitutable, the information distortion of the observed demand from real demand becomes more serious (Widiarta et al., 2008b). Any excess demand for an item A would be satisfied by a substitutable item B, and this phenomenon is invisible to the inventory manager. Thus, the inventory manager is apt to under-forecast the demand of item A, but over-forecast the demand of item B (Widiarta et al., 2008b).

Widiarta et al. (2008b) contended that top-down forecasting is less affected by the information distortion, because top-down forecasting uses historical demand at group level, which could be less dependent upon the degree of item substitutability. The South Korean Navy purchased a series of weapon systems from the same manufacturers to ensure stability of supply and continued technical support. For instance, the South Korean Navy installed MTU (Motoren- und Turbinen-Union) series as a main diesel engine for all the large warships: MTU 20V 956 TB92 for Destroyers; MTU 16V 538 TB82 for Frigates; and MTU 12V 956 TB82 for Corvettes (Saunders, 2009). This allows a large proportion of the spare parts to be substitutable. While substitutability causes the difficulty of forecasting spare parts, this can make the performance of top-down forecasting better than direct forecasting.

Maintenance for warships in the South Korean Navy is classified as Military maintenance or contract maintenance (Korean Navy, 2003). Military maintenance refers to a maintenance conducted by the Military maintenance units themselves. Contract maintenance refers to a maintenance conducted by contractors with the Military. When Military maintenance is impossible, the contract maintenance is to be conducted. Contract maintenance could induce non-normality. Information on spare parts consumed by the Military maintenance is updated on the logistical database in the Naval Logistics Command; however, information on spare parts consumed by contract maintenance is likely to be missing from the logistical database. Therefore, the entire records of the spare parts demand could not be stored on the logistical database.

The information distortion caused by the above sources provides unreliable data to the inventory manager for demand forecasting, then an inaccurate demand forecast caused by the faulty data introduces large stock holdings (Choi et al., 2005, Lee, 2007). As stated above, the first and the second sources are unavoidable because these originate from the true demand fluctuations. However, the other three sources of the non-normal demands would generate proxy demand data because they distort true demand. The reliability of the data input system should be established to prevent "garbage in garbage out". Forecasts based on the proxy demands are likely to be biased. This might be a limitation of this research. True demand data for the purpose of repair and maintenance, not for the purpose of stocking, should be used for generating forecasts. The entire information on the spare parts demand, including demand for the contract maintenance and substitution, should be included in the logistical database, however, these data are not currently captured.

It should be noted that research found that top-down forecasting could outperform direct forecasting for missing or unreliable data as stated in Subsection 2.5.2 (Schwarzkopf et al., 1988). This suggested that hierarchical forecasting could be more accurate than direct forecasting for the spare parts data obtained from the South Korean Navy.

4.6 Summary and Conclusion

This research particularised a specific group of warships and time boundary for data collection. The time boundary was decided from January 2002 to November 2007, and three types of warships were selected; and eight pieces of equipment installed in these warships were selected. Then, 300 items were chosen. The 300 items were ranked in terms of historical dollar volume within the same type of equipment and the same NATO Supply Classification Group (NSCG). Finally, two nearest items in terms of historical dollar volume formed a group.

The demand features of the spare parts were analysed by several measures. The time series were found to be non-normal and correlate within a group. Thus, the findings in the literature (Markland, 1970, Businger and Read, 1999, Eaves and Kingsman, 2004) that military spare parts demands are non-normal were repeated in this research.

Comparing the features of group level time series with the features of item level time series, Cv(size), Pr(peak) and Pr(zero) decreased at group level time series. This identified reduced non-normal demand features at group level time series than those at item level time series. This reduced non-normality of group level time series suggests that hierarchical forecasting would be superior to direct forecasting (Gross and Sohl, 1990, Fliedner and Lawrence, 1995, Fliedner, 2001).

In most of the data, seasonality was non-significant, although some monthly seasonal effects were significant. Long procurement lead time was identified in the spare parts. Long procurement lead time together with long review cycle introduces large stock holdings. However, this long procurement lead time is a feature which makes hierarchical forecasting more accurate than direct forecasting (Shlifer and Wolff, 1979).

Some relative demand features which are different in each equipment group were identified. Gun/RD was characterised as having higher intermittency, smaller demand volume, shorter lead time, and more expensive prices. ME was characterised as having lower correlation, steeper downward trend, lower intermittency, and larger demand volume. GE/AC was characterised as having higher variability, greater peakedness, and greater deviation from a normal distribution.

In order to implement simple exponential smoothing in the next chapter for both direct and hierarchical forecasting methods, trend component and seasonal component have to be measured or removed (Waters, 1991, Gardner Jr. and Diaz-Saiz, 2002). Thus, linear trend and additive seasonality adjustment was implemented. The adjustment removed trends from the time series; however, outliers were still prominent after the adjustment.

Five sources of non-normality of data were identified: a) a few large customers; b) operating factors; c) the multi-echelon inventory systems; d) the budgeting process; and

e) the maintenance system. The former two sources reflect true demand fluctuations; however, the latter three sources would generate rather proxy demands than true demands. This might be a limitation of this research.

However, for these missing or unreliable data hierarchical forecasting could present better performance than direct forecasting (Schwarzkopf et al., 1988). With respect to the source e) the maintenance system, it was identified that a large proportion of the Naval spare parts is substitutable. This substitutability of the spare parts is a feature which makes hierarchical forecasting more accurate than direct forecasting (Widiarta et al., 2008b).

As the first chapter of findings, this chapter found the nature of the demand to be non-normal. Data features which could make hierarchical forecasting more accurate than direct forecasting were also identified. Therefore, this chapter answers research question a) "what is the nature of the spare parts demand in the South Korean Navy?" Then, this chapter created the appropriate conditions to produce hierarchical forecasting; that is, data selection, grouping and trend and seasonal adjustment. Based on this chapter, both direct and hierarchical forecasting methods can be produced and the performance of the forecasting methods are compared in the following chapters.

Chapter 5. Forecasting The Spare Parts Demand

In Chapter 4, the nature of the demand for warship spare parts in the South Korean Navy was shown to be non-normal. Demand features, which could make hierarchical forecasting more accurate than direct forecasting, were also identified. These include long forecasting horizons (Shlifer and Wolff, 1979), a reduction of non-normality at group level time series (Gross and Sohl, 1990, Fliedner and Lawrence, 1995, Fliedner, 2001), substitutability in many spare parts (Widiarta et al., 2008b), the hierarchical demand structure using the types of equipment and the National Stock Number (NSN), and the unreliability of data (Schwarzkopf et al., 1988). In this chapter, a range of direct and hierarchical forecasting methods are compared.

As stated in the beginning of Chapter 2, forecasting performance is situational (DeLurgio, 1998, Chatfield, 2004). The special property of intermittent data creates a particular difficulty in selecting an appropriate accuracy measure (Syntetos and Boylan, 2005). As shown in Chapter 4, the South Korean Navy holds a large stock of spare parts to cover the non-normal demands from warships for long lead time periods. The performance of the Naval inventory systems for spare parts crucially influenced by its forecasting performance could affect the operational availability of the weapon systems using those spare parts. This requires a well-identified forecasting method that accurately predicts the spare parts demand.

Three groups of accuracy measures such as absolute, relative and derivative measures were reviewed in Section 2.6. The absolute and relative measures do not capture the monetary value and the service level of each item and so do not present the practical impact of a forecasting method upon the inventory system. Derivative measures have been suggested as a practical resolution to verify forecasting performance (Wemmerlöv, 1989, Sani and Kingsman, 1997, Heuts et al., 1999, Eaves, 2002). The performance of forecasting methods was measured by the three groups of measures in this research.

This chapter begins with a review of the forecasting procedure in the South Korean Navy in Section 5.1. The absolute and relative measures are clarified in Section 5.2. The derivative measure is clarified in Section 5.3. Forecasting processes and results under direct and hierarchical forecasting strategies are presented in Section 5.4 and 5.5 respectively. Finally, a summary and concluding remarks are presented in Section 5.6.

5.1 Forecasting Spare Parts Demand in the South Korean Navy

An annual demand (AD) denotes an estimate of an annual demand for an item for the next year. An annual demand is a point forecast consisting of a single number. The South Korean Navy does not produce a prediction interval consisting of upper and lower limits. The Navy forecasts annual demand for items in the Authorised Stock item List (ASL). An annual demand is forecast on the basis of yearly aggregated time series over the most recent five years expressed as five observations (Seon and U, 2009). The Navy uses three simple univariate demand forecast techniques (i.e. naïve average; simple moving average; and least square method) for forecasting annual demands (Seon and U, 2009).

Naïve average assumes that on examining historical data from period to period the changes are not significant enough to be taken into consideration (Waller, 2003). Naïve average can be defined as in equation (5-1). Naïve average can smooth out random fluctuations. It is appropriate when the observed time series have no noticeable trend and seasonality (Hyndman et al., 1998). Simple moving average was reviewed in Subsection 2.2.3.

$$\hat{y}_t(1) = \frac{1}{n}\sum_{i=1}^{n} y_{t-i+1} \tag{5-1}$$

where:

$\hat{y}_t(1)$ = the one step ahead forecast made at time period t

y_t = the observed demand for an item at time period t

n = the number of the total time periods

A linear trend [equation (4-7)] in a time series can be estimated by fitting a straight line. Assuming the trend will continue to future values, extrapolation using the fitted trend is possible. Least square method assumes that the relationship between the dependent variable, $\breve{y}_t$, and the independent variable, t, can be approximated by a straight line (Bowerman et al., 2005, Silver et al., 1998). Least square method determines the best straight line for the given time series by minimising equation (4-8) (Waller, 2003, Cryer and Chan, 2008). Then, least square method can generate forecasts using equation (5-2).

$$\hat{y}_t(\tau) = \hat{\beta}_0 + \hat{\beta}_1(t+\tau) \tag{5-2}$$

where:

$\hat{y}_t(\tau)$ = the τ steps ahead forecast made at time period t

$\hat{\beta}_0$ and $\hat{\beta}_1$ = the estimated parameters

The confidence limits can be calculated as $\hat{y}_t(\tau) \pm z_{\frac{1-p}{2}} \times \sqrt{VAR(\tau)}$ where p is the probability value for the confidence limits, $z_{\frac{1-p}{2}}$ is the upper $100(\frac{1-p}{2})\%$ point from the standard normal distribution, and $Var(\tau)$ is the τ steps ahead forecast variance. For example, a 95% confidence limits of the forecast can be calculated as $\hat{y}_t(\tau) \pm 1.96 \times \sqrt{VAR(\tau)}$ (McCleary, 1980). However, the South Korean Navy does not produce confidence limits.

Requirement objective (RO1) at wholesale-level and requisition objective (RO2) at retail-level are determined based upon the annual demand. The South Korean Navy provides criteria for the selection of a forecasting method: a) naive average is used for the items in which the continuing annual demands are approximately similar or irregular; b) simple moving average is used for items which exhibit downward trends; and c) least square method is used for items which exhibit upward trends (Seon and U, 2009). The inventory manager in the Navy selects a forecasting method for an item based on the criteria and his experience (Seon and U, 2009). However, the Navy provides no specific criteria expressed as the value of a demand feature or equations.

As shown in Subsection 1.2.1 the forecasting accuracy in terms of volume has been very low. Generally this low accuracy might be caused by the nature of the demand characterised as non-normal which was identified in Chapter 4. In addition, the current forecasting methods in the Navy are too naïve to catch the characteristics of the demand so the methods have failed to generate accurate forecasts (Choi et al., 2005). When a time series deviates from a normal distribution, least square method will be inaccurate (Miles and Shevlin, 2001). Moving average could not produce an accurate forecast for spare parts demand compared with other methods. For instance, some authors argued that the performance of moving average for forecasting spare parts demand for the UK Air Force (Eaves and Kingsman, 2004) and vehicles and agricultural machinery (Sani and Kingsman, 1997) was inferior by Croston's method and Syntetos-Boylan Approximation. Naïve average was found to be superior for forecasting the spare parts demand in the UK Air Force to moving average; however, this was inferior to exponential smoothing, Croston's method and Syntetos-Boylan Approximation (Eaves and Kingsman, 2004). More accurate direct forecasting methods for spare parts than the three forecasting methods currently used in the South Korean Navy were identified in Subsection 2.2.3.

As stated in Section 1.3, the comparisons of various direct forecasting methods were beyond the scope of this research. The main objective of this research was to compare the alternative forecasting strategies within the context of spare parts demand for the South Korean Navy. Clarifying reliable accuracy measures might be an important precondition for comparing the forecasting strategies.

5.2 Absolute and Relative Measures of Accuracy

The intermittent and erratic characteristics of the spare parts demand require a deliberate choice of accuracy measures. Three groups of measures were identified. This section clarifies absolute and relative measures to be used for this research. Derivative measures are described in the next section.

As stated in Subsection 2.6.1, it is appropriate to use mean absolute deviation (MAD) in order to avoid heavier weight on large errors; root mean square error (RMSE) can be also useful when large errors cause greater inventory costs in proportion to small errors. In order to capture these two alternative effects of large errors upon weight (or costs), both MAD [equation (2-44)] and RMSE [equation (2-41)] were employed as absolute measures.

In order to compare relative forecasting errors across a large amount of data, accuracy measures should be standardised. Thus, as stated in Subsection 2.6.1, error measures divided by means were introduced by researchers (Regattieri et al., 2005, Boylan et al., 2008). MAD/A [equation (2-45)] and RMSE/A [equation (2-46)] were employed for this research.

The mean rank of each forecasting method also provides a rough measure of forecasting performance across the overall data set (Kling and Bessler, 1985, Sani and Kingsman, 1997). The forecasting performance of each forecasting method was measured and ranked in terms of MAD and RMSE, and then the ranks of each forecasting method were averaged across the overall data series.

A parametric test is defined as "a statistical test which assumes that scores used come from a population of scores which is normally distributed" (Howitt and Cramer, 2008, p. 515). A non-parametric test is defined as "a statistical test of significance which requires fewer assumptions about the distribution of values in a sample than a parametric test" (Howitt and Cramer, 2008, p. 514). A statistical method for comparing the means of several groups is analysis of variance (ANOVA) (Moore et al., 2009). Although an assumption of ANOVA is that the data are from a normally distributed population, ANOVA F-test is robust to the non-normality (Neter et al., 1990). The spare parts demand data were found to be extremely non-normally distributed as shown in Subsection 4.3.4. In this case, previous research (Sani and Kingsman, 1997, Eaves, 2002) on comparing the performance of forecasting methods for spare parts demand, that were non-normal, used Friedman's non-parametric test. This research used Friedman's test based on the previous research comparing forecasting methods for spare parts demand.

If there are more than two conditions (defined as forecasting methods) and the same cases (defined as spare parts) are used in all conditions, Friedman's non-parametric test can be used in order to test differences between experimental conditions (Friedman, 1937). The mean rank of each forecasting method can be examined with Friedman's test, because Friedman's test is based on ordinal data. With a sum of ranks for each condition (i.e. each forecasting method), the test statistic of Friedman's test, *Fr* was calibrated as in equation (5-3) (Kanji, 2006). The test statistic, *Fr* has a tendency to be distributed according to χ^2 distribution (see Table A-1) with k-1 degrees of freedom (Howitt and Cramer, 2008).

$$Fr = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1) \qquad (5\text{-}3)$$

where:

R_i = the sum of ranks for each forecasting method

n = the total sample size

k = the number of conditions (i.e. forecasting methods)

As mentioned in Subsection 2.6.1, cumulative or running sum of forecast errors (RSFE) [equation (2-48)] can detect the biased tendency of a forecasting method (Narasimhan et al., 1998). While unbiased forecasts lead RSFE to near zero, over-forecasting leads to negative RSFE, and under-forecasting leads to positive RSFE. A relative size of RSFE, S was calculated as RSFE divided by MAD [equation (2-49)].

For the purpose of comparing the two alternative forecasting strategies (i.e. hierarchical forecasting versus direct forecasting), an unbiased relative measure was used [equation (2-57)] (Dangerfield and Morris, 1992). As a summary statistic, a natural log of the ratios is defined as in equation (5-4), equation (5-5), or equation (5-6). A positive log relative error indicates that a direct forecasting method is superior to a hierarchical forecasting method, and vice versa. LN(ratio) or LN(HF/DF) denotes this natural log relative error in this research. For example, LN(HF/*tsm*) for MAD indicates $\ln(MAD_{HF}/MAD_{tsm})$.

$$Log\ relative\ error(MAD) = \ln(MAD_{HF} / MAD_{DF}) \tag{5-4}$$

$$Log\ relative\ error(RMSE) = \ln(RMSE_{HF} / RMSE_{DF}) \tag{5-5}$$

$$Log\ relative\ error(Inventory\ costs) = \ln(Inventory\ costs_{HF} / Inventory\ costs_{DF}) \tag{5-6}$$

5.3 Derivative Measures of Accuracy

Derivative measures use simulation to derive the impact of forecasting accuracy in terms of inventory levels and service levels achieved by the inventory system. This section formulates the simulation model used for the derivative measures. The simulation process is illustrated with a case of an item. Then, the processes of measuring simulation results and model selection are clarified.

5.3.1 Simulation model structure

The simulation model employed in this research is similar to the implied stock-holding model which was reviewed in Chapter 2 (Eaves, 2002, Eaves and Kingsman, 2004). The simulation of this research is based on the wholesale-level procurement system which is described in Chapter 4. This model is characterised as the (R, S) control system; that is, the periodic-review, order-up-to-level system (Silver et al., 1998).

The real historical demand and procurement lead time data from January 2002 to November 2007, which were employed for generating forecasts and measuring the performance of the forecasts, were used for this simulation. For simplicity, two restrictions of the simulation model against the real case were imposed. First, the retail-level replenishment systems were not considered. This was because the purpose of this simulation was to verify the performance of forecasting methods using the inventory system and not to verify the performance of the inventory system in itself. Second, the procurement decision was only allowed once a year (i.e. in every January of the years considered), although in the real life case a modification of the annual procurement decision is allowed for an exceptional case.

Figure 5-1 illustrates the simulated inventory system. The simulation was a deterministic process. A particular demand forecast for an item always produced the same inventory position. The simulation was based on discrete events. Each demand was assumed to occur on the first day of every month from January 2005 to November 2007. The procurement decisions were made at three times (i.e. January 2005, January 2006 and January 2007). The order-up-to-level (R) was calculated until November 2007 in order to avoid a biasing effect caused by redundant inventory after the simulation periods. Thus, the last order quantities in January 2007 were computed as the amounts of expected demand until November 2007. Forecast quantity can influence the inventory position after the first delivery's arrival which are ordered by an order-up-to-level (S) based on the first forecasting. The performance was measured in each month after the first deliveries arrived, because forecasting quantities are unable to have any effect on the simulation performance before the first deliveries arrive (Eaves, 2002).

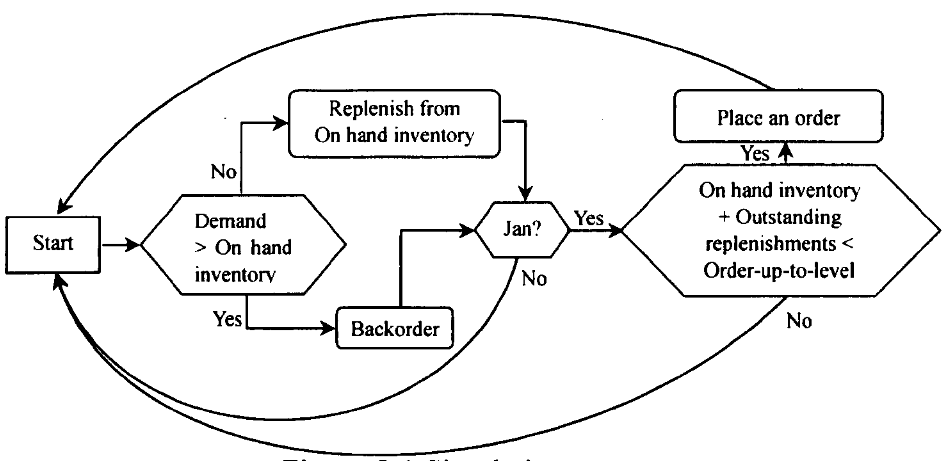


Figure 5-1 Simulation process

The simulation was conducted with the assumptions shown in Table 5-1. A primary packaged quantity indicates a packaged quantity in which a unit of the product is delivered. The primary packaged quantity for all spare parts was assumed as one unit. Fractions were rounded up to the nearest integer.

Table 5-1 Assumptions for the simulation

a)	Simulation period = monthly, January 2005 ~ November 2007
b)	Measuring period = monthly, the first delivery arrival ~ November 2007
c)	Initial inventory position = inventory consumed before the first delivery arrival
d)	Review interval (R) = twelve months procurement decision in January
e)	Order-up-to-level (S) = lead time demand + review period + safety level
f)	Stock on hand open at time t = stock on hand close at time t-1
g)	Order quantity = S – (stock on hand open + delivery + stock on order)
h)	Primary packaged quantity = one unit

A lead time demand refers to an estimated demand during the lead time. Eaves (2002) set the initial inventory position as an equal amount to the first lead time demand plus one unit. However, as different forecasting methods produce different lead time demands, different initial stocks by the different forecasting methods for the same item have a possibility of causing a biasing effect upon the simulation's performance. Therefore, in this research, the initial inventory position of an item was established as the same amount of inventory for all the forecasting methods; that is, the inventory consumed after January 2005 and before the first delivery arrival. As the future demand is invisible, establishing the initial inventory position as the inventory consumed before the first delivery might be unrealistic. However, this might be fair for all the forecasting methods because all the forecasting methods were given the same initial inventory position. Too high an initial inventory position which cannot be consumed until the end of the simulation periods would be unfairly advantageous to under-forecasting. Too low an initial inventory position which will be depleted from the start of the simulation periods would also be unfairly advantageous to over-forecasting. Therefore, establishing the initial inventory position as the same amount to the inventory consumed before the first delivery arrival would have no biasing effect upon the comparisons of the forecasting methods which can be introduced by the initial inventory position based on different forecasts as in Eaves (2002).

5.3.2 Simulation process

Table 5-2 exemplifies the simulation process. This simulation employs a forecasting method, namely a simple combination (SC) between an exponential smoothing model with quarterly aggregated data adjusted for linear trend at group level and an exponential smoothing model with monthly aggregated data adjusted for additive seasonality at item level (*SCtqsm*). Detailed description of forecasting methods will be provided later in this chapter. Using *SCtqsm*, three forecasts were generated for the simulation; that is, forecasts in January 2005, January 2006 and January 2007. The initial inventory position was set as 17 units because the historical demand before the first delivery arrival (i.e. in this case until September 2005) is 17 units. Delivery was assumed to be completed at the end of the procurement lead time (PROLT), then to be used from the first day of the next month. Assuming the PROLT of the exemplified item is 9 months, the first delivery would arrive on the last day of September 2005, and the items can be replenished from the first day of October 2005. A stock on hand close was calculated as the sum of stock on hand open and delivery quantity minus demand quantity. The stock on hand close in October 2005 was calculated as "16 = 0 + 17 – 1".

Table 5-2 Example of the simulation process

	Demand Qty	Stock On Hand Open	Stock On Order	Order Qty	Delivery Qty	Stock On Hand Close	Stock Qty	Back Order Qty	Unfilled Demand Qty	Forecast Qty	Safety Level	Order-up-to-Level
Jan-05	5	17		17		12	12			29.67	4.24	34
Feb-05		12	17			12	12					
Mar-05	6	12	17			6	6					
Apr-05		6	17			6	6					
May-05		6	17			6	6					
Jun-05	1	6	17			5	5					
Jul-05	1	5	17			4	4					
Aug-05	4	4	17									
Sep-05		0	17									
Oct-05	1	0			17	16	16					
Nov-05	4	16				12	12					
Dec-05	1	12				11	11					
Jan-06	2	11		20		9	9			26.38	3.77	31
Feb-06	4	9	20			5	5					
Mar-06	8	5	20			-3		3	3			
Apr-06		-3	20			-3		3				
May-06		-3	20			-3		3				
Jun-06	1	-3	20			-4		4	1			
Jul-06		-4	20			-4		4				
Aug-06		-4	20			-4		4				
Sep-06	2	-4	20			-6		6	2			
Oct-06	1	-6			20	13	13					
Nov-06		13				13	13					
Dec-06		13				13	13					
Jan-07		13		4		13	13			14.31	1.95	17
Feb-07		13	4			13	13					
Mar-07	2	13	4			11	11					
Apr-07		11	4			11	11					
May-07	1	11	4			10	10					
Jun-07		10	4			10	10					
Jul-07	2	10	4			8	8					
Aug-07	4	8	4			4	4					
Sep-07		4	4			4	4					
Oct-07	1	4			4	7	7					
Nov-07	1	7				6	6					

A safety level was calculated by multiplying the forecast quantity (i.e. lead time demand and review period) by the safety period (i.e. three months) divided by the sum of the PROLT and review period. The first forecast quantity, the safety period, the PROLT, and the review period are 29.67 units, 3 months, 9 months, and 12 months respectively. Thus, the first safety level was calculated as a quantity for three months; that is, "4.24 units = 29.67 × [3 / (9 + 12)]". The first order-up-to-level (S) was calculated as the safety level plus forecast quantity (i.e. lead time demand and review period); that is, "33.91 units = 4.24 + 29.67". Then, the first order-up-to-level is rounded up to 34 units which are multiple of the primary packaged quantity (i.e. one unit). The last safety level was calculated as a quantity for one and a half months which is a half quantity of the safety level in the previous periods. This is because the demand data were only available until November 2007 (i.e. 11 months from the last order point, which is roughly a half of the forecasting horizon of the two previous periods). If the last safety level is established as the same quantity as the previous safety level (i.e. the quantity for 3 months), this would introduce a biasing effect which can be advantageous to under-forecasting. Therefore, the last safety level was calculated by multiplying the forecast quantity (i.e. estimated demand for eleven months) by the safety period (i.e. one and a half months) divided by forecasting periods (i.e. eleven months); that is, "1.95 units = 14.31 × (1.5 / 11)".

An order quantity was calibrated as order-up-to-level (S) minus the sum of stock on hand open, delivery and stock on order. The first order quantity was calculated as "17 = 34 – (17 + 0 + 0)". The first backorder occurred in March 2006. The demand quantity, 8 units, is 3 units larger than stock on hand open, 5 units, without any delivery arrival. Thus, 3 units were backordered until the next replenishment arrival which was planned in October 2006. These backorders accumulate continuously until October 2006. Thus, the maximum stock-out reached 6 units in September 2006.

5.3.3 Measures of inventory model performance

As stated in Subsection 2.7.3, the performance of an inventory model from a simulation can be measured by the total inventory costs and the inventory fill rate (Petrovica et al., 1998). It was also mentioned that the total inventory costs can be calculated as the sum of inventory carrying costs and inventory stock-out costs (Sterman, 1989, Sani and Kingsman, 1997, Petrovica et al., 1998, An et al., 2002). As noted in Subsection 2.7.3, quantifying inventory stock-out costs is difficult, especially in militaries (MacDonald, 1997).

Total inventory costs

For the purpose of sorting out the difficulty in quantifying stock-out costs, two kinds of approach, namely safety margin approach and total inventory costs approach, were employed in previous research. The safety margin approach refers to a simulation method which avoids the quantification of stock-out costs (Wemmerlöv, 1989, Eaves, 2002, Eaves and Kingsman, 2004). The total inventory costs approach refers to a simulation method which attempts to establish a reasonable criterion to quantify stock-out costs (Sterman, 1989, Sani and Kingsman, 1997, An et al., 2002).

The safety margin approach used by some authors (Wemmerlöv, 1989, Eaves, 2002, Eaves and Kingsman, 2004), which provides spare parts to user units as a fill rate 100% by adding the safety stock levels to the order-up-to-level until no stock-outs occurred, were reviewed in Subsection 2.7.3. This safety margin approach could be considered to be an appropriate methodology to a simulation for a military inventory system. This is because militaries are unlikely to reconcile themselves to the risk of stock-out. However, assuming an inventory system which holds the exact same amount of safety stock to achieve a 100% fill rate is rather unrealistic. In practice, the safety stock level is determined by a forecast for the demand and cannot be adjusted to the exact amount of no stock-out because the future demand for the safety stock is invisible. The shortage of some items of spare parts is unavoidable in militaries. As mentioned in Chapter 1, although the US Military continues to hold as much as 60% excess stock, inventory shortages for some items still occur (Hinton Jr., 1999). Therefore, a reasonable criterion to quantify stock-out costs for spare parts was required.

As mentioned above, the total inventory costs can be calculated as the sum of inventory carrying costs and inventory stock-out costs. However, it is difficult to quantify the inventory costs, especially the inventory stock-out costs. Waller (2003) suggested a rough stock-out costs structure which consists of 55% of a backorder case, 25% of a lost sale case, and 20% of a lost customer case. For example, if the costs of a backorder, a lost sale, and a lost customer are £5, £50 and £500, respectively, the expected stock-out costs would be: $0.55\times£5 + 0.25\times£50 + 0.20\times£500 = £115.25$.

In the South Korean Navy, every unfilled spare part demand is backordered. Then, the unfilled spare parts are either expedited for a special case with a specially allocated budget or supplied in the next regular delivery. The case of lost customers can be thought of as being equivalent to non-operational weapon systems caused by either incomplete repair or incomplete maintenance. As such, the costs of lost customers also need to be considered in the stock-out costs.

As stated in Subsection 2.7.3, previous researchers (Sterman, 1989, Sani and Kingsman, 1997, Petrovica et al., 1998, An et al., 2002) quantified the total inventory costs by their own criteria. When the information of unit variable costs is available, quantifying the total inventory costs as a fraction of unit variable cost might be rational as with the total inventory costs structure by Sani and Kingsman (1997).

Excess stock holdings increase inventory carrying costs (Waller, 2003). However, stock-out costs can dramatically outweigh unit variable cost as stated in Section 4.1. Avoiding a stock-out situation in order to sustain operational availability of weapon systems and a fast response to the demand might be more important than saving inventory carrying costs.

Thus, a stock-out case in the South Korean Navy might have to be charged more penalty costs than a stock-out case in business; for example, 33% of the item costs in the above case of Sani and Kingsman (1997). As with the costs structure of An et al. (2002), weighing twice the inventory carrying costs on the inventory stock-out costs might be appropriate. As such, in this research, following An et al. (2002), the total inventory costs were calculated using equation (5-7).

$$\begin{aligned} \textit{Total inventory costs} = \textit{unit variable cost} \times (&\textit{mean inventory per month} \\ &\times 0.2 + \textit{mean stock-out per month} \times 0.4) \end{aligned} \qquad (5\text{-}7)$$

In this research, the performance of a forecasting method was measured after the first delivery arrival until November 2007. In the exemplified simulation process, the performance was measured between October 2005 and November 2007. This is because the first delivery arrived on the last day of September 2005. The unit variable cost of the item, the monthly mean inventory position, and the monthly mean backorder quantity are ₩828,353 (£423), 6.73 units, and 1.04 units respectively. Using equation (5-6), the total inventory costs were calculated as "₩1,549,558 (£792) = $(6.73 \times 0.2 + 1.04 \times 0.4) \times$ ₩828,353".

Inventory fill rate

The performance of a simulation model can also be measured by an inventory fill rate. This research used equation (2-62) to compute the inventory fill rates. In the exemplified simulation process, the performance was measured between October 2005 and November 2007 (i.e. 26 months). The monthly mean shortage was calculated as the sum of unfilled demand quantity divided by simulation periods; that is, "$0.23 = 6/26$". The monthly mean demand between October 2005 and November 2007 was 1.35 units. As the fill rate is to be calculated as one minus the mean shortage divided by the mean demand, it was calculated as "$0.83 = 1 - (0.23/1.35)$".

The inventory fill rate is easier to calculate than the total inventory costs. However, the inventory fill rate is incapable of summarising the total performance of a model. By simply adding inventories to the safety level regardless of increasing inventory carrying costs, the inventory fill rate can be increased. Moreover, the inventory fill rate could not capture the volume and monetary value of all the items, so that the inventory fill rate cannot summarise the total performance appropriately. However, the total inventory costs can comprise the carrying inventory quantities and the stock-out quantities as well as the monetary values of all the items, although they are difficult to calculate exactly. Therefore, in this research, the total inventory costs were employed as the major performance measure, whereas the inventory fill rate was employed as a qualifying measure. The South Korean Navy considers that a fill rate for the spare parts, greater than 70%, is to be an acceptable fill rate. Therefore, for this research, a fill rate greater than 70% was assumed to be qualified as an acceptable fill rate.

5.3.4 Simulation model selection

The two mentioned approaches toward the total inventory costs were compared. Table 5-3 presents the results of two simulation models. Model I employed the total inventory costs approach; model II employed the safety margin approach. As the simulation model of this research, the process of model I was exemplified earlier in this chapter. As described in Chapter 2, in model II, the safety stock level increased iteratively to the order-up-to-level until no stock-outs occurred under the (R, S) control system (Eaves, 2002, Eaves and Kingsman, 2004). Real data for the selected 300 items were used for this simulation. Simple exponential smoothing using a software package, Forecast 2.04 (Hyndman, 2010), was employed to generate forecasts. For the purpose of comparing the simulation models, the two models used the same forecasting quantities.

Table 5-3 Simulation model comparisons

	Model I		Model II		S
	Inventory costs	Mean fill rate	Inventory costs	Mean fill rate	
um	₩778,942,252	0.865	₩3,709,522,615	1.00	-129.79
tm	₩804,463,179	0.630	₩3,733,496,873	1.00	91.91
sm	₩887,983,370	0.869	₩4,183,624,548	1.00	-136.46
tsm	₩790,657,985	0.762	₩3,910,508,135	1.00	-8.3
uq	₩816,285,294	0.859	₩3,859,006,687	1.00	-133.74
tq	₩804,120,116	0.623	₩3,687,836,343	1.00	98.14
sq	₩981,537,203	0.851	₩4,599,250,682	1.00	-129.81
tsq	₩922,928,237	0.679	₩4,312,433,923	1.00	37.54
uy	₩1,368,692,175	0.880	₩6,175,031,932	1.00	-147.37
ty	₩1,185,235,950	0.760	₩3,927,614,229	1.00	24.82

S = the relative size of RSFE: a negative S value indicates that the forecasting method continuously generated over-forecasts, and vice versa.

Positive *S* values of *tm, tq, tsq* and *ty* denoted that these forecasting methods continuously generate under-forecasts. Negative *S* values of *um*, *sm*, *tsm*, *uq*, *sq*, and *uy* denoted that these forecasting models continuously generate over-forecasts. In Model I, *um* (followed by *tsm*) minimised the inventory costs. *tm*, *tq*, and *tsq* were disqualified by their fill rates of smaller than 70%. On the other hand, in Model II, *tq* (followed by *um*) minimised the inventory costs. Comparing the total inventory costs of the forecasting methods for the two simulation models, Model II was found to be advantageous to under-forecasts. Figure 5-2 describes the change of the ranks for the forecasting methods in terms of the total inventory costs in the two simulation models. As shown in the figure, the under-forecasts have a tendency of being ranked higher in Model II than Model I in terms of the total inventory costs; the over-forecasts have a tendency of being ranked lower in Model II than Model I.

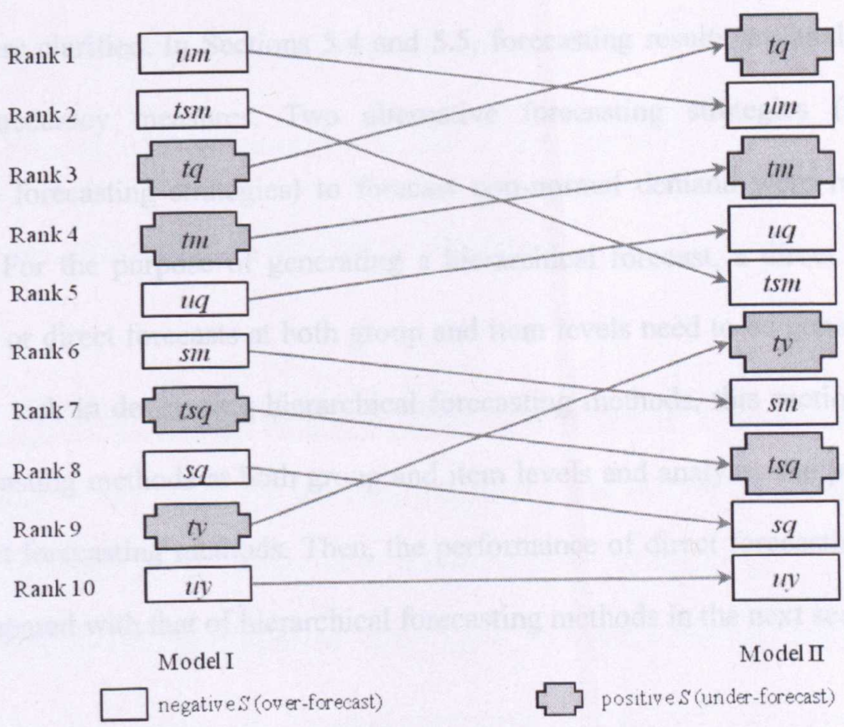


Figure 5-2 Forecasting performance comparisons in the two simulation models

This advantageous position to under-forecasts in Model II might be caused by the unrealistic adding of the safety stocks. By adding safety stock up to the exact amount of no stock-out, under-forecasts can minimize unnecessary stocks, so that under-forecasts score lower total inventory costs in Model II. As a matter of fact, however, the safety stock level is determined by a forecast quantity and cannot be adjusted to the exact amount of no stock-out, because the future demand of the safety stock is invisible. As mentioned, the shortage of some items of spare parts is unavoidable in militaries. Model I was employed as the simulation model for this research, because Model I establishes the safety stock level according to a forecast quantity, so it is closer to a real life situation.

5.4 Direct Forecasting

In the previous sections, the three accuracy measures which were employed for this research were clarified. In Sections 5.4 and 5.5, forecasting results are analysed using the three accuracy measures. Two alternative forecasting strategies (direct and hierarchical forecasting strategies) to forecast non-normal demand were reviewed in Chapter 2. For the purpose of generating a hierarchical forecast, a direct forecast at group level or direct forecasts at both group and item levels need to be generated. As a preliminary task in developing hierarchical forecasting methods, this section develops direct forecasting methods at both group and item levels and analyses the performance of the direct forecasting methods. Then, the performance of direct forecasting methods will be compared with that of hierarchical forecasting methods in the next section.

5.4.1 The development of direct forecasting methods

A variety of direct forecasting methods for non-normal demand were reviewed in Subsection 2.2.3. Some authors (Businger and Read, 1999, Jiafu et al., 2009) used the Box-Jenkins models to forecast spare parts demand. As mentioned in Section 3.7, the Box-Jenkins models were the initial models used to examine the spare parts demand data obtained from the South Korean Navy. This research used a series of diagnostic tests (for the Box-Jenkins models) such as the analysis of residuals against fitted values, the analysis of the standardized residuals, and the analysis of probability plots. The new residuals ($\hat{\varepsilon}_t$) and the standardized residuals (e_t) can be expressed as in equation (5-8) and (5-9) respectively.

$$\hat{\varepsilon}_t = y_t - \hat{y}_t \tag{5-8}$$

$$e_t = (y_t - \hat{y}_t) / \sqrt{\hat{p}_t} \tag{5-9}$$

where: y_t = the observed demand for an item at time period t

$\hat{y}_t$ = the estimated demand for the item at time period t

$\hat{p}_t$= the estimated error variance at time period t

In equation (5-8), the new residuals, $\hat{\varepsilon}_t$, against the fitted values, $\hat{y}_t$, should give a random scatter if the model fits the data (Shumway and Stoffer, 2006). However, in most of the 300 units of data, there were still some patterns remaining in the plots after the Box-Jenkins models had fitted the data. Figure 5-3 exemplifies the plots of residuals, $\hat{\varepsilon}_t$, against the fitted values, $\hat{y}_t$. Monthly aggregated time series for four spare parts (i.e. Starting Valve, Cylinder Ring, Air Reduction Valve, and Flexible Shaft for ME II) were fitted with ARMA (1, 1), ARMA (2, 1), ARMA (1, 1) and ARIMA (1, 1, 0) which were found to be the best fitting models for the time series respectively. However, as shown in the figure, all the plots failed to give a random scatter.

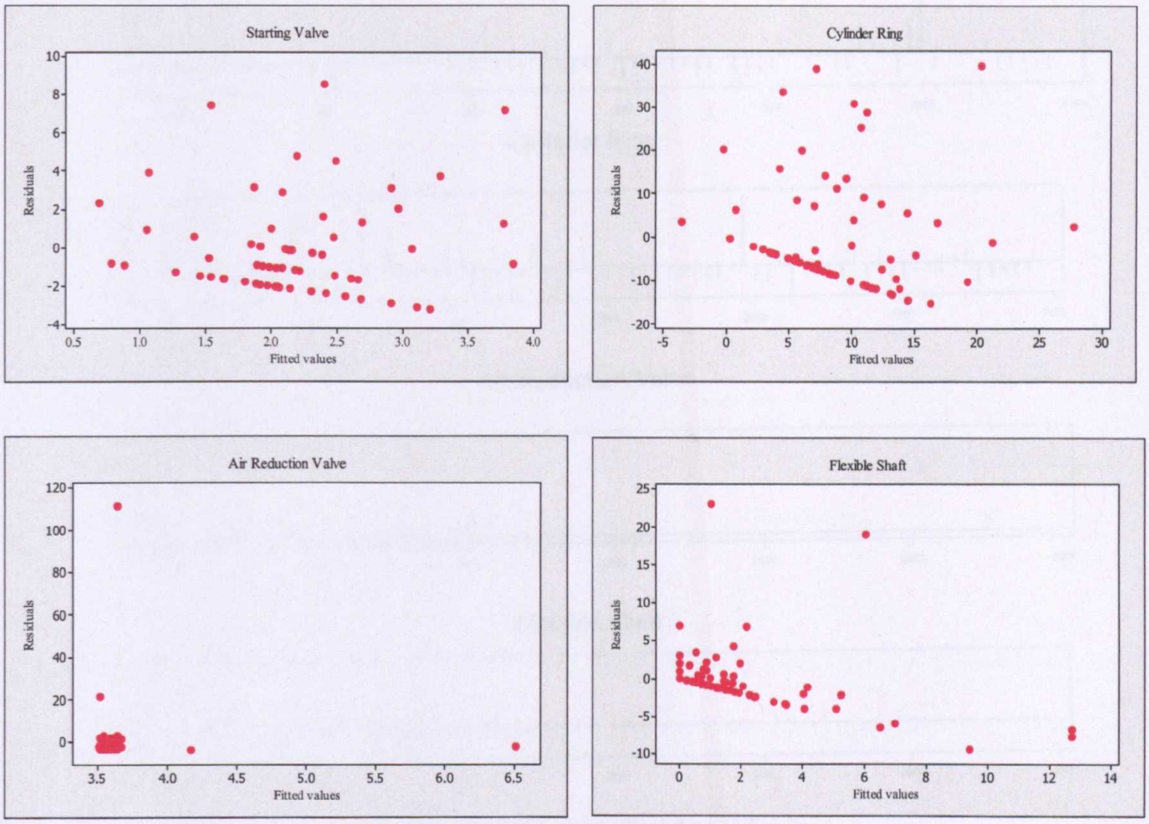


Figure 5-3 Residuals against fitted values

In equation (5-9), the standardized residuals, e_t, should behave as an independent and identically distributed (iid) sequence with a mean of 0 and variance of 1 if the model fits the data (Shumway and Stoffer, 2006). Figure 5-4 presents the standardized residuals plots for the above four spare parts. As shown in the figure, all the plots failed to provide an iid sequence. Assuming standardized residuals have a normal distribution 99% of the standardized residuals should lie between -2.58 and +2.58. All the residuals violated this assumption as 95.45% of residuals for Starting Valve, 95.77% of those for

Cylinder Ring, 98.59% of those for Air Reduction Valve, and 97.18% of those for Flexible Shaft lay between -2.58 and +2.58.

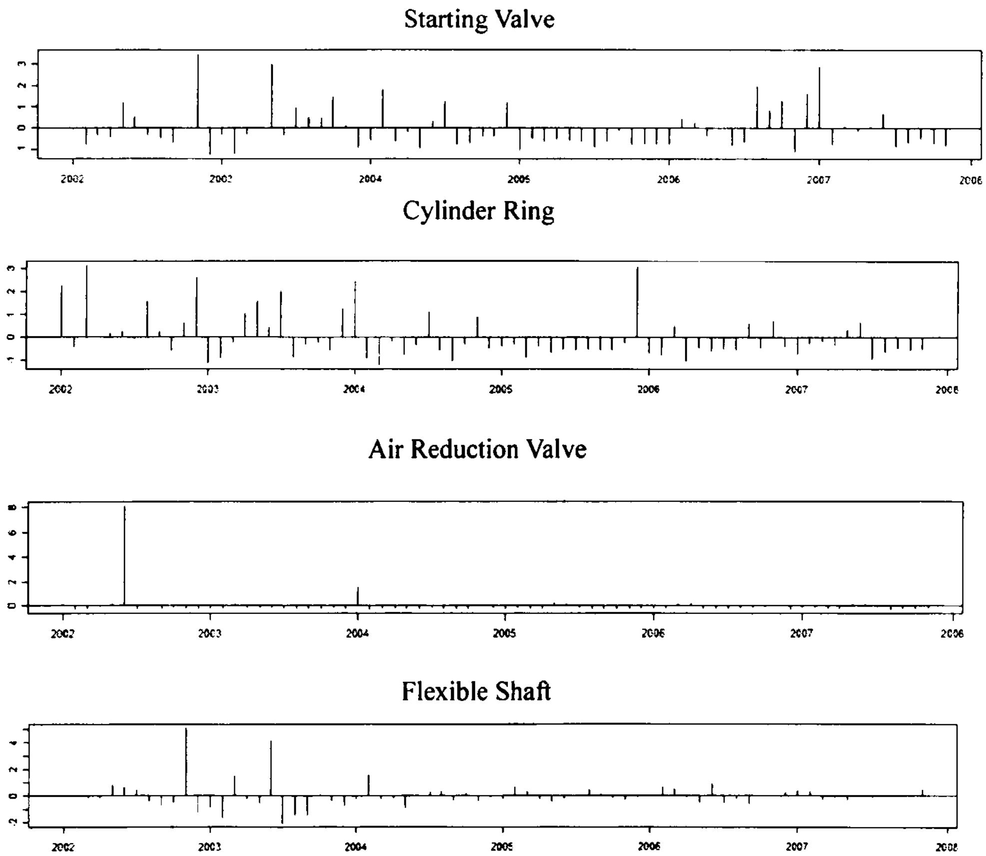


Figure 5-4 Standardized residuals

Figure 5-5 presents probability plots against a normal distribution for the above four spare parts. The distributions of the data were non-normal and skewed (as the curves deviated from a straight line and were asymmetrical). Outliers, where the line jumps to the end, were identified. Based on the above diagnostic tests, the Box-Jenkins models were found to not fit the spare parts demand data obtained from the South Korean Navy.

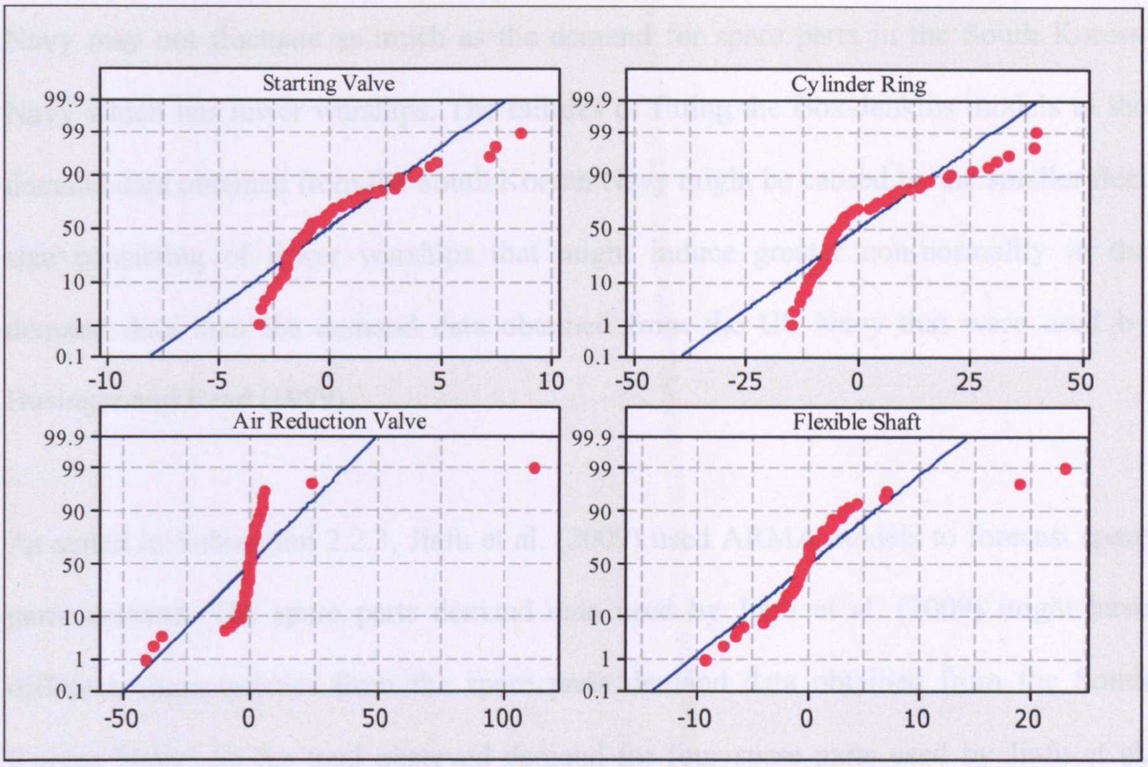


Figure 5-5 Probability plots of the residuals

As stated in Subsection 2.2.2, although complex multivariate models could generate a better fit to given data than univariate models, this supremacy does not necessarily convert into superior forecasts due to the sensitivity of multivariate models to changes in structure (Chatfield, 2004). Multivariate models can be expected to give more accurate forecasts than univariate models when a high cross-correlation between time series, such as financial time series, is observed. However, Willemain et al. (1994) found that the sparseness of the empirical intermittent data makes most of the cross-correlation non-significant. Multivariate models are unlikely to produce more accurate forecasts for intermittent demand than univariate models.

As stated in Subsection 4.5.2, a few large customers (i.e. warships) could induce non-normality (Silver, 1970, Eaves, 2002). Considering the large number of warships in the US Navy (Saunders, 2009), it can be expected that the demand for spare parts in the US

Navy may not fluctuate as much as the demand for spare parts in the South Korean Navy which has fewer warships. The failures of fitting the Box-Jenkins models to the demand data obtained from the South Korean Navy might be caused by the smaller fleet size consisting of fewer warships that might induce greater non-normality in the demand data than the demand data obtained from the US Navy that were used by Businger and Read (1999).

As stated in Subsection 2.2.3, Jiafu et al. (2009) used ARMA models to forecast spare parts demand. The spare parts demand data used by Jiafu et al. (2009) might have different characteristics from the spare parts demand data obtained from the South Korean Navy. As the total observed demand for four spare parts used by Jiafu et al. (2009) in 2007 was 12,665 units, the mean yearly demand per item can be calculated as 3,166.25 units. As shown in Table 4-6, the mean yearly demand per item of this research was 189.83 units. The spare parts data from the South Korean Navy were probably more intermittent and fluctuated more than the data used by Jiafu et al. (2009). Large volumes of demand might have led to reasonable performance of ARMA models in the investigation of Jiafu et al. (2009).

As stated in Subsection 2.2.3, the standard forecasting method for non-normal demand is exponential smoothing (Croston, 1972, Sani and Kingsman, 1997, Narasimhan et al., 1998). As stated in Subsection 2.5.2, simple exponential smoothing has been shown to be superior to moving average or other complex models when used with hierarchical forecasting (Fliedner and Lawrence, 1995, Fliedner, 1999, Dekker et al., 2004, Viswanathan et al., 2008). The suitability of simple exponential smoothing for hierarchical forecasting was discussed in Subsection 2.5.2 (Fliedner, 1999). As stated in

Section 1.3, the main objective of this research was comparing direct and hierarchical forecasting strategies. In the assessment of forecasting strategies for spare parts in the South Korean Navy this research uses simple exponential smoothing as a benchmark to compare the alternative forecasting strategies. In this research, simple exponential smoothing was adopted as an individual forecasting method for both forecasting strategies (i.e. hierarchical forecasting and direct forecasting) due to the following reasons:

a) the performance of simple exponential smoothing has been well-verified in research (Fliedner and Lawrence, 1995, Fliedner, 1999, Dekker et al., 2004, Viswanathan et al., 2008);
b) simple exponential smoothing is suitable for the purpose of forecasting spare parts demand for controlling spare parts inventory (Fliedner, 1999);
c) simple exponential smoothing is suitable for forecasting a large number of spare parts which is the case with spare parts for warships in the South Korean Navy, because simple exponential smoothing requires less time to generate forecasts especially with a large number of items (Fliedner, 1999);
d) the potential trend and seasonality of spare parts demand can be removed through trend and seasonal adjustment, so that simple exponential smoothing can be used for forecasting spare parts demand.

Simple exponential smoothing (SES) for forecasting spare parts was generated utilising a forecasting software-package, i.e. "Forecast 2.04" (Hyndman, 2010). As mentioned in Subsection 2.5.2, the smoothing parameters might not have a significant effect on the relative forecasting performance of hierarchical and direct forecasting methods

(Dangerfield and Morris, 1992). Thus, optimal smoothing parameters which were generated in "Forecast 2.04" have been simply employed.

As a holistic single case design (Yin, 2003), forecasts were generated at the wholesale-level; that is, the Naval Logistics Command (NLC). The NLC decides annual demand (AD) for items in Authorised Stock item List (ASL) once a year, although modification is allowed throughout a year. Likewise, forecasts for this research were produced at four times: i.e. January 2004, January 2005, January 2006, and January 2007. These forecasting results were assumed to be used as a basis of annual procurement decisions in this research. Forecasting horizons were established as the period including procurement lead time (between 3 and 18 months) and review period (12 months). Therefore, forecasting horizons were set up as quite long periods between 15 months and 30 months (i.e. procurement lead time + review period). As stated in Subsection 4.3.2, the period of data obtained from the Navy was January 2002 to November 2007. Hence, the forecasts were generated until November 2007; the performance of the forecasts was measured until November 2007.

Each forecast was based on all the previous periods. For example, the forecast for 2005 (or 2006) used data for the periods between January of 2002 and December of 2004 (or 2005). Based on monthly aggregated data, SES generates a one period ahead (i.e. one month ahead) forecast. In order to generate a long period forecast, this one period ahead forecast was multiplied by the forecasting horizons. For example, in order to generate a 20 months ahead forecast, the one period ahead forecast was multiplied by 20. Based on quarterly or yearly aggregated data, a one period ahead forecast was multiplied by the fraction of the quarters or the years. For example, in order to generate a 20 months ahead forecast, the one period ahead forecast based on quarterly aggregated data was multiplied by 20/3.

As shown in Section 4.4, the data were adjusted for the four data adjustment methods (*u, t, s*, and *ts*) to generate the forecasts. After generating the forecasts using the adjusted data, the forecasts were reverse adjusted so that the forecasts could be compared with actual demands.

5.4.2 Group level forecasting

As mentioned in Section 1.3, item level forecasting is the major concern of this research. However, group level forecasting should also be investigated because the group level forecasting accuracy might affect item level forecasting accuracy with a proration method (Gross and Sohl, 1990). Simple exponential smoothing forecasting models were generated for predicting the group level demand of the 150 pair groups which were clarified in Subsection 4.3.2. Forecasting results for four different measuring periods were provided, that is, 2004 ~ 2007, 2005 ~ 2007, 2006 ~ 2007, and 2007. For example, for the forecasting measuring period, 2004 ~ 2007, four forecasts were generated for a pair group using each forecasting method in January of 2004, January of 2005, January of 2006 and January of 2007. Likewise, for the forecasting measuring period, 2005 ~ 2007, three forecasts were generated for a pair group using each forecasting method in January of 2005, January of 2006 and January of 2007. As such, the forecasting errors were measured once a year.

Table 5-4 provides a summary of the performance of direct forecasting methods for the 150 pair group level demand as the sums of the standardised errors. As stated in Section

5.2, for the purpose of comparing forecasting performance across a large amount of data, forecasting errors (MAD and RMSE) were divided by the mean demand of the time series (Regattieri et al., 2005, Boylan et al., 2008). MAD and RMSE for each forecast have been divided by the yearly mean demand of the data which were used for generating the forecast. For example, MAD and RMSE between 2005 and 2007 were divided by the yearly mean demand of the data used for generating the 3 years of forecasts between 2002 and 2004. As such, MAD/A and RMSE/A for each forecast were generated. Then, *S* (RSFE/MAD), MAD/A, and RMSE/A for the 150 pair groups were summed up respectively to present the summary of the performance for each forecasting method. Table 5-4 presents these sums of the standardised errors for each forecasting method.

Table 5-4 Sums of direct forecasting errors at group level

	Year	*um*	*tm*	*sm*	*tsm*	*um*	*tq*	*sq*	*tsq*	*uy*	*ty*
S	04 ~ 07	-92.40	-28.82	-97.67	-59.50	-97.20	-21.45	-95.55	-32.20	-81.22	**-2.32**
	05 ~ 07	-68.42	59.12	-74.23	**17.50**	-82.42	60.10	-79.29	30.03	-66.30	28.09
	06 ~ 07	**-34.97**	89.56	-49.29	53.09	-55.76	91.14	-54.63	67.39	-43.50	80.69
	07	-32.00	58.00	-38.00	**22.00**	-54.00	66.00	-52.00	44.00	-58.00	60.00
MAD /A	04 ~ 07	92.11	208.84	107.18	125.86	96.52	133.72	109.56	124.63	96.17	**78.58**
	05 ~ 07	86.01	176.33	86.27	**75.81**	93.44	97.05	94.43	88.73	90.44	93.65
	06 ~ 07	79.68	194.88	77.05	**75.21**	86.15	104.41	82.33	87.75	85.95	95.48
	07	54.50	172.54	53.65	**51.31**	56.94	69.05	56.06	60.18	59.07	67.72
RMSE /A	04 ~ 07	109.89	256.39	134.86	177.01	113.54	175.67	135.31	167.01	112.66	**93.29**
	05 ~ 07	99.84	191.48	100.98	**90.58**	108.84	112.20	111.67	104.79	104.42	110.15
	06 ~ 07	89.20	207.51	86.26	**85.33**	97.57	116.66	93.27	98.61	95.80	106.51
	07	54.50	172.54	53.65	**51.31**	56.94	69.05	56.06	60.18	59.07	67.72

The balanced forecasts in terms of *S* and the forecasting methods giving minimmum MAD/A or RMSE/A are shown in bold.

The forecasting results in 2004 ~ 2007 and others were obviously different. This might be caused by the high influence of the two peak points in June 2002 and March 2003 previously mentioned upon the forecasts in 2004. A relative size of RSFE, *S* (RSFE/MAD) denotes the tendency of over- and under-forecasts. As stated in Subsection 2.6.1, *S* should be used with caution, because the purpose of *S* is to monitor the balance of negative and positive errors. An S close to zero does not mean an accurate forecast, but a balanced forecast. In the years 2004 ~ 2007, all forecasting methods produced an over-forecast due to the high influence of peak points in 2002 and 2003 upon the forecasts in 2004. As stated in Section 2.4, Gross and Sohl (1990) contended that the applicability of top-down forecasting when a consistent direction of group level forecasting is observed. The consistent over-forecasting caused by the peak points might have an implication for a suitable condition of hierarchical forecasting.

With the exception of the forecasts in 2004 ~ 2007, the forecast with data adjusted for linear trend (*t*) and the forecast with data adjusted for linear trend and additive seasonality (*ts*) were under-forecasts, whereas the forecast with unadjusted data (*u*) and the forecast with data adjusted for additive seasonality (*s*) were over-forecasts.

MAD/A and RMSE/A showed similar results. In the years 2004 ~ 2007, the forecast with yearly aggregated data adjusted for linear trend (*ty*) minimised MAD/A and RMSE/A. However, with the exception of the year 2004, the forecast with monthly aggregated data adjusted for linear trend and additive seasonality (*tsm*) followed by the forecasts with monthly aggregated unadjusted data (*um*) dominated by minimum MAD/A and RMSE/A.

The above results were corroborated as shown in Table 5-5. Table 5-5 presents the mean ranks of direct forecasting methods at group level. Friedman's non-parametric test was conducted. The number of treatments is 10 (i.e. the number of direct forecasting methods) and the number of blocks is 150 (i.e. the number of the groups). In the years 2004 ~ 2007, the forecast with yearly aggregated data adjusted for linear trend (*ty*) minimised the mean rank in terms of MAD and RMSE; whereas, with the exception of the forecasts in 2004 the forecast with monthly aggregated data adjusted for linear trend and additive seasonality (*tsm*) followed by the forecast with monthly aggregated unadjusted data (*um*) minimised the mean rank in terms of MAD and RMSE.

Table 5-5 Mean ranks of direct forecasting methods at group level
(Friedman's test for forecasting errors: p-value < 0.001)

	Year	*um*	*tm*	*sm*	*tsm*	*uq*	*tq*	*sq*	*tsq*	*uy*	*ty*
MAD	04 ~ 07	4.04	8.03	5.54	6.70	4.30	7.45	5.42	6.54	3.95	**3.01**
	05 ~ 07	4.95	*6.21*	5.19	**3.90**	5.93	*6.37*	5.89	5.33	5.24	5.95
	06 ~ 07	4.73	*6.51*	4.91	**4.21**	5.35	*7.01*	5.41	5.16	5.35	*6.16*
	07	4.95	*5.97*	5.21	**4.63**	5.38	*6.39*	5.35	5.25	5.48	*6.05*
RMSE	04 ~ 07	3.87	8.01	5.25	7.33	4.29	7.36	5.32	6.76	3.74	**3.07**
	05 ~ 07	4.98	6.01	5.21	**3.87**	6.07	6.31	6.24	5.31	5.01	5.96
	06 ~ 07	4.70	6.53	4.93	**4.33**	5.38	7.01	5.47	5.23	5.24	5.97
	07	4.95	5.97	5.21	**4.63**	5.38	6.39	5.35	5.25	5.48	6.05

The mean rank of the forecasting method among the 10 direct forecasting methods over the 150 groups in terms of MAD or RMSE is presented in each cell; the forecasting methods minimising the mean rank in terms of MAD or RMSE are shown in bold.

Table 5-6 Forecasting accuracy comparisons by data aggregation (group)

	Year	*m*	*q*	*y*
MAD	04 ~ 07	6.08	5.93	**3.48**
	05 ~ 07	**5.06**	5.88	5.59
	06 ~ 07	**5.09**	5.73	5.75
	07	**5.19**	5.59	5.77
RMSE	04 ~ 07	6.12	5.93	**3.40**
	05 ~ 07	**5.02**	5.98	5.49
	06 ~ 07	**5.13**	5.77	5.61
	07	**5.19**	5.59	5.77

The mean rank of the forecasting methods using the data aggregation method in Table 5-5 is presented in each cell; the data aggregation methods minimising the mean rank in terms of MAD or RMSE are shown in bold.

Table 5-6 presents a summary of Table 5-5 in terms of the data aggregation method. In the years 2004 ~ 2007, the forecast with yearly aggregated data (*y*) minimised the mean rank in terms of MAD and RMSE; however, with the exception of the forecasts in 2004, the forecast with monthly aggregated data (*m*) minimised the mean rank in terms of

MAD and RMSE.

Table 5-7 Forecasting accuracy comparisons by data adjustment (group)

	Year	*u*	*t*	*s*	*ts*
MAD	04 ~ 07	**4.10**	6.19	5.35	6.52
	05 ~ 07	5.37	6.15	5.49	**4.96**
	06 ~ 07	5.14	6.41	5.20	**4.85**
	07	5.27	5.90	5.52	**4.97**
RMSE	04 ~ 07	**3.96**	6.19	5.23	6.84
	05 ~ 07	5.35	6.20	5.52	**4.91**
	06 ~ 07	5.11	6.48	5.15	**4.89**
	07	5.27	5.90	5.52	**4.97**

The mean rank of the forecasting methods using the data adjustment method in Table 5-5 is presented in each cell; the data adjustment methods minimising the mean rank in terms of MAD or RMSE are shown in bold.

Table 5-7 presents a summary of Table 5-5 in terms of the data adjustment method. The forecast with unadjusted data (*u*) presented the minimum mean rank in the years 2004 ~ 2007, and the second minimum mean rank in other periods. The forecast with data adjusted for linear trend and additive seasonality (*ts*) minimised the mean rank with the exception of the year 2004.

5.4.3 Item level forecasting

Item level forecasting was the major subject of this research. Simple exponential smoothing forecasting models were generated for the 300 items. As with group level forecasting, forecasting results for four different measuring periods were provided, that is, 2004 ~ 2007, 2005 ~ 2007, 2006 ~ 2007, and 2007. For example, three forecasts were generated for an item using each forecasting method in January of 2005, January of 2006 and January of 2007. The results of the item level forecasting were similar to the above results of the group level forecasting. Table 5-8 presents the results of the item level direct forecasting over the 300 items. As in Table 5-4, this table also presents the sum of the standardised errors for each forecasting method. In the years 2004 ~ 2007, every forecasting method had a tendency towards over-forecasting. From the year 2005, there was a tendency of under-forecast in the forecast with data adjusted for linear trend (*t*) and the forecast with data adjusted for linear trend and additive seasonality (*ts*) with the exception of the forecast with monthly aggregated data adjusted for linear trend and additive seasonality (*tsm*) in the years 2005 ~ 2007.

Table 5-8 Sums of forecasting errors (item)

	Year	*um*	*tm*	*sm*	*tsm*	*uq*	*tq*	*sq*	*tsq*	*uy*	*ty*
S	04 ~ 07	-160.50	-44.61	-175.89	-114.31	-157.24	**-31.66**	-155.68	-56.13	-173.30	-32.21
	05 ~ 07	-129.79	91.91	-139.46	**-8.30**	-133.74	98.14	-129.81	37.54	-147.37	24.81
	06 ~ 07	-84.46	147.80	-101.85	**54.85**	-88.53	148.52	-96.60	91.64	-100.26	121.17
	07	-82.00	106.00	-88.00	**14.00**	-80.00	116.00	-98.00	54.00	-104.00	92.00
MAD /A	04 ~ 07	**258.36**	375.24	293.09	352.99	262.04	362.35	297.85	347.15	381.59	284.77
	05 ~ 07	214.59	240.43	221.64	**209.91**	226.08	243.02	234.05	231.04	309.00	282.18
	06 ~ 07	214.75	254.65	210.83	**208.39**	221.28	255.52	220.34	227.56	292.26	271.73
	07	147.76	167.05	147.60	**143.05**	149.58	168.58	151.08	150.78	221.49	186.66
RMSE /A	04 ~ 07	314.13	498.58	365.96	477.92	**311.65**	473.30	370.84	455.47	439.47	347.74
	05 ~ 07	**249.19**	283.93	259.08	250.98	262.67	286.49	276.68	275.72	350.52	338.24
	06 ~ 07	237.15	284.22	**232.63**	233.10	246.06	286.10	245.95	254.89	319.11	307.39
	07	147.76	167.05	147.60	**143.05**	149.58	168.58	151.08	150.78	221.49	186.66

The balanced forecasts in terms of *S* and the forecasting methods giving minimum MAD/A or RMSE/A are shown in bold.

MAD and RMSE were divided by the yearly mean demand of the data which had been used for generating the forecast. With the exception of 2004, *tsm* presented minimum MAD/A, which is consistent with the forecasting result at group level. However, the forecasting method giving minimum RMSE/A was different from the result for group level forecast in each row. Moreover, the item level forecast giving minimum MAD/A in each row was incongruous with the item level forecast giving minimum RMSE/A in each row.

Compatible results were observed in terms of the mean ranks of direct forecasting methods at item level, as shown in Table 5-9. This supports the result of group level forecasting. The number of treatments is 10 (i.e. the number of direct forecasting methods) and the number of blocks is 300 (i.e. the number of the selected items). In the years 2004 ~ 2007, the forecast with yearly aggregated data adjusted for linear trend (*ty*) minimised the mean rank in terms of MAD and RMSE. However, the forecast with monthly aggregated data adjusted for linear trend and additive seasonality (*tsm*) dominated by minimising the mean rank when the forecasting performance in 2004 was ruled out. The forecast with monthly aggregated unadjusted data (*um*) followed the performance of *tsm* in terms of the mean rank.

Table 5-9 Mean ranks of direct forecasting methods at item level
(Friedman's test for forecasting errors: p-value < 0.001)

	Year	*um*	*tm*	*sm*	*tsm*	*uq*	*tq*	*sq*	*tsq*	*uy*	*ty*
MAD	04 ~ 07	4.08	7.76	5.39	6.66	4.36	7.28	5.30	6.39	4.24	**3.53**
	05 ~ 07	4.84	6.16	5.22	**4.50**	5.53	6.28	5.76	5.42	5.24	6.02
	06 ~ 07	4.76	6.36	4.99	**4.55**	5.28	6.59	5.40	5.15	5.47	6.27
	07	4.96	5.85	5.41	**4.71**	5.42	6.20	5.63	5.23	5.58	5.64
RMSE	04 ~ 07	4.00	7.79	5.29	7.09	4.26	7.24	5.17	6.58	4.03	**3.53**
	05 ~ 07	4.87	6.15	5.13	**4.43**	5.59	6.29	5.91	5.38	5.07	6.15
	06 ~ 07	4.70	6.43	4.87	**4.58**	5.19	6.62	5.43	5.20	5.41	6.40
	07	4.96	5.85	5.41	**4.71**	5.42	6.20	5.63	5.23	5.58	5.64

The mean rank of the forecasting method among the 10 direct forecasting methods over the 300 items in terms of MAD or RMSE is presented in each cell; the forecasting methods minimising the mean rank in terms of MAD or RMSE are shown in bold.

Table 5-10 Forecasting accuracy comparisons by aggregation (item)

	Year	*m*	*q*	*y*
MAD	04 ~ 07	5.97	5.83	**3.88**
	05 ~ 07	**5.18**	5.74	5.63
	06 ~ 07	**5.17**	5.61	5.87
	07	**5.24**	5.62	5.61
RMSE	04 ~ 07	6.05	5.81	**3.78**
	05 ~ 07	**5.14**	5.79	5.61
	06 ~ 07	**5.14**	5.61	5.90
	07	**5.24**	5.62	5.61

The mean rank of the forecasting methods using the data aggregation method in Table 5-9 is presented in each cell; the data aggregation methods minimising the mean rank in terms of MAD or RMSE are shown in bold.

Table 5-10 presents a summary of Table 5-9 in terms of data aggregation methods. Table 5-11 presents a summary of Table 5-9 in terms of data adjustment methods. Table 5-10 and Table 5-11 corroborate the result of the group level forecasting. In the years 2004 ~ 2007, the forecast with yearly aggregated data (*y*) and the forecast with unadjusted data (*u*) minimised the mean rank in terms of MAD and RMSE; however, with the exception of 2004, the forecast with monthly aggregated data (*m*) and the forecast with data adjusted for linear trend and additive seasonality (*ts*) minimised the mean rank in terms of MAD and RMSE. The forecast with unadjusted data (*u*) presented the second minimum mean ranks with the exception of the year 2004.

Table 5-11 Forecasting accuracy comparisons by data adjustment (item)

	Year	*u*	*t*	*s*	*ts*
MAD	04 ~ 07	**4.23**	6.19	5.35	6.52
	05 ~ 07	5.20	6.15	5.49	**4.96**
	06 ~ 07	5.17	6.41	5.20	**4.85**
	07	5.32	5.90	5.52	**4.97**
RMSE	04 ~ 07	**4.09**	6.19	5.23	6.84
	05 ~ 07	5.17	6.20	5.52	**4.91**
	06 ~ 07	5.10	6.48	5.15	**4.89**
	07	5.32	5.90	5.52	**4.97**

The mean rank of the forecasting methods using the data adjustment method in Table 5-9 is presented in each cell; the data adjustment methods minimising the mean rank in terms of MAD or RMSE are shown in bold are shown in bold.

Additionally, the forecasting results were compared in terms of the equipment groups identified in Subsection 4.3.4 so as to test whether there is any preference of forecasting methods for a particular group of equipment. For these comparisons, the forecasts in 2004 were ruled out because there was an obvious biasing effect of the two peak points on the forecasting performance. At the outset, forecasting results of a 4 group scheme (Gun, ME, GE/AC, and RD) in 2005 ~ 2007 were investigated. However, due to the small group size (i.e. only 6 items) of RD, a significant result for RD could not be found. Thus, these groups were required to be combined. The reasonableness of combining the equipment groups was presented in terms of mean demand, and the functions and the links of the pieces of equipment in Subsection 4.3.4. It was shown that Gun and RD can be combined into a group. In addition to the reasonableness of the combination, RD and Gun represented the similar preference for forecasting methods (i.e. *um* minimised the mean rank in terms of MAD and RMSE in both Gun and RD). Therefore, RD was integrated with Gun. As such, Friedman's test was conducted.

Table 5-12 Mean ranks of direct forecasting methods for equipment groups at group level (Friedman's test for forecasting errors, p-values are shown in the 3rd column)

	Group	p-value	No.	*um*	*tm*	*sm*	*tsm*	*uq*	*tq*	*sq*	*tsq*	*uy*	*ty*
MAD	Gun/RD	$p = 0.075$	22	4.2	6.5	5.9	4.4	4.8	6.7	5.3	5.7	5.7	6.0
	ME	$p < 0.001$	94	4.9	6.7	4.9	**4.0**	5.7	6.7	5.5	5.6	5.0	5.8
	GE/AC	$p < 0.001$	34	5.4	4.7	5.6	**3.4**	7.2	5.2	7.3	4.4	5.5	6.2
RMSE	Gun/RD	$p = 0.104$	22	4.3	6.3	5.7	4.6	5.0	6.8	5.7	5.5	4.8	6.3
	ME	$p < 0.001$	94	5.1	6.3	4.9	**3.8**	6.0	6.5	5.9	5.5	5.1	5.9
	GE/AC	$p < 0.001$	34	5.1	5.1	5.8	**3.5**	7.0	5.5	7.4	4.7	5.0	5.9

No. indicates the number of the pair groups within the equipment group; the mean rank of the forecasting method over the pair groups within the equipment group in terms of MAD or RMSE is presented in each cell; the forecasting methods minimising the mean rank in terms of MAD or RMSE are shown in bold.

Table 5-12 presents the mean ranks of direct forecasting methods for equipment groups at group level. The number of treatments is 10 (i.e. the number of direct forecasting methods) and the numbers of blocks (i.e. the number of item pairs in each equipment group) are 22 for Gun/RD, 94 for ME, and 34 for GE/AC. Gun/RD was non-significant as the p-value was greater than 0.05. *tsm* minimised the mean rank in terms of MAD and RMSE for ME and GE/AC. The results of ME and GE/AC were consistent with the results of the overall group level forecasting.

Table 5-13 presents the mean ranks of direct forecasting methods for equipment groups at item level. At item level forecasting, *um* minimised the mean rank for Gun/RD; *tsm* minimised the mean rank for ME and GE/AC. This is consistent with the group level forecasting results.

Table 5-13 Mean ranks of direct forecasting methods for equipment groups at item level (Friedman's test for forecasting errors, p-values are shown in the 3rd column)

	Group	p-value	No.	*um*	*tm*	*sm*	*tsm*	*uq*	*tq*	*sq*	*tsq*	*uy*	*ty*
MAD	Gun/RD	$p = 0.001$	44	**3.7**	6.1	5.2	5.4	4.6	6.5	5.4	6.1	5.9	6.0
	ME	$p < 0.001$	188	5.0	6.5	5.2	**4.5**	5.5	6.4	5.6	5.5	5.0	5.9
	GE/AC	$p < 0.001$	68	5.1	5.4	5.4	**3.9**	6.3	5.8	6.5	4.8	5.5	6.4
RMSE	Gun/RD	$p = 0.001$	44	**3.8**	6.1	5.3	5.4	4.6	6.6	5.3	6.2	5.8	6.0
	ME	$p < 0.001$	188	5.1	6.3	5.1	**4.4**	5.5	6.4	5.8	5.4	4.9	6.1
	GE/AC	$p < 0.001$	68	5.0	5.7	5.2	**4.0**	6.3	5.8	6.7	4.7	5.1	6.4

No. indicates the number of the items within the equipment group; the mean rank of the forecasting method over the items within the equipment group in terms of MAD or RMSE is presented in each cell; the forecasting methods minimising the mean rank in terms of MAD and RMSE are shown in bold.

Recalling the average statistics for the 3 equipment groups in Subsection 4.3.4, there were some differences in data features among the equipment groups. Gun/RD was characterised as having higher intermittency, smaller demand volume, shorter lead time, and more expensive price. ME was characterised as having lower correlation, steeper downward trend, lower intermittency, and larger demand volume. GE/AC was characterised as having higher variability, greater peakedness, and greater deviation from a normal distribution.

Some of the differences among the 3 equipment groups might be a reason for the different performance of the forecasting methods for different equipment groups. Based on these results of the different forecasting performance upon different equipment groups, it starts to suggest the consideration of different forecasting methods for different equipment groups.

In summary, there were some inconsistencies in the results of direct forecasting method in terms of S, MAD/A, RMSE/A. However, some obvious patterns were found in the mean ranks of the forecasting methods. At both group and item levels, the forecast with monthly aggregated data adjusted for linear trend and additive seasonality (*tsm*) was the most robust direct forecasting method between 2005 and 2007. Some differences in the performance of direct forecasting methods for the three equipment groups were also identified.

5.4.4 Simulation

So far, the performance of direct forecasting methods has been measured with absolute measures. However, the absolute measures do not measure the practical impact that a forecasting method has on the inventory system. It has been suggested that derivative measures are more practical (Wemmerlöv, 1989, Sani and Kingsman, 1997, Heuts et al., 1999, Eaves, 2002). This approach uses simulation to derive the impact of forecasting accuracy in terms of the inventory levels and the service levels achieved by the inventory system. Corresponding with the above results of direct forecasting methods, simulations, which were clarified earlier, were conducted for every month from January 2005 to November 2007 with the same data which had produced the forecasts. Then, the simulation results were compared in terms of the total inventory costs and the inventory fill rate.

Table 5-14 provides simulation results for the direct forecasting using simple exponential smoothing. The ranks (MAD and RMSE) are based on Table 5-9. *um* minimised the total inventory costs, although it ranked as 2nd in terms of MAD and RMSE; *tsm* ranked as 2nd in terms of the total inventory costs in spite of its 1st rank in terms of MAD and RMSE. The fill rates of *um* and *tsm* (i.e. 86.5% and 76.2% respectively) were considered to be acceptable fill rates. *tm*, *tq*, and *tsq*, were disqualified, because the fill rates of these were lower than 70%.

Table 5-14 Simulation results for direct forecasting

Total	Total inventory costs	Rank	Mean stock	Mean backorder	Fill rate	Rank (MAD)	Rank (RMSE)
um	₩778,942,252.81	1	79373.50	1410.71	0.865	2	2
tm	₩804,463,179.71	4	40606.17	16337.82	0.630	9	8
sm	₩887,983,370.44	6	70100.92	1942.35	0.869	3	4
tsm	₩790,657,985.15	2	40981.07	7644.91	0.762	1	1
uq	₩816,285,294.67	5	81554.87	1547.44	0.859	6	6
tq	₩804,120,116.83	3	35784.89	16699.36	0.623	10	10
sq	₩981,537,203.85	8	74358.24	2470.49	0.851	7	7
tsq	₩922,928,237.58	7	35267.36	12607.54	0.679	5	5
uy	₩1,368,692,175.98	10	79308.20	1390.90	0.880	4	3
ty	₩1,185,235,950.28	9	48538.46	9227.46	0.760	8	9

Rank (MAD) or Rank (RMSE) indicates the rank of the forecasting methods in terms of MAD or RMSE between 2005 and 2007.

The performance of the direct forecasting methods in terms of the total inventory costs and the mean rank according to the inventory costs of the forecasting methods are presented in Table 5-15.

Table 5-15 Inventory costs and mean ranks of direct forecasting

	Total inventory costs		Friedman's test ($p < 0.001$)	
	Costs	Rank	Mean rank	Rank
um	₩778,942,252.81	1	5.14	3
tm	₩804,463,179.71	4	5.24	5
sm	₩887,983,370.44	6	5.82	8
tsm	₩790,657,985.15	2	4.54	1
uq	₩816,285,294.67	5	6.01	9
tq	₩804,120,116.83	3	5.19	4
sq	₩981,537,203.85	8	6.36	10
tsq	₩922,928,237.58	7	5.05	2
uy	₩1,368,692,175.98	10	5.35	6
ty	₩1,185,235,950.28	9	5.59	7

The mean rank of the forecasting method among the 10 direct forecasting methods over the 300 items in terms of the total inventory costs is presented in the 4th column.

Friedman's test for the inventory costs was conducted. The number of treatments is 10 and the number of blocks is 300. In contrast to the results of the total inventory costs in

Table 5-14, *tsm* minimised the mean rank. This performance of *tsm* was consistent with the results of MAD and RMSE. The 1st ranked forecast in terms of the total inventory costs (i.e. *um*) was just found as the 3rd ranked forecast in terms of the mean rank.

Table 5-16 presents the simulation results for the direct forecasting methods sorted by the three equipment groups. *um* minimised the inventory costs for Gun/RD; *tsm* minimised the inventory costs for ME and GE/AC. Although *tm* and *tq* for GE/AC presented lower total inventory costs than those of *tsm*, both *tm* and *tq* were disqualified by their fill rates lower than 70%. These results were consistent with the results from the absolute measures (i.e. MAD and RMSE) in Table 5-13.

Table 5-16 Simulation results of direct forecasting for the equipment groups

	Gun/RD		ME		GE/AC	
	Inventory costs	Fill rate	Inventory costs	Fill rate	Inventory costs	Fill rate
um	**₩115,040,970**	0.89	₩601,551,261	0.85	₩62,012,736	0.89
tm	₩172,273,596	0.57	₩578,268,916	0.63	₩53,536,033	0.66
sm	₩154,750,262	0.94	₩668,430,661	0.84	₩64,501,528	0.91
tsm	₩164,695,241	0.73	**₩569,675,266**	0.77	**₩55,921,397**	0.77
uq	₩133,818,259	0.89	₩618,702,964	0.84	₩63,332,090	0.89
tq	₩178,429,279	0.58	₩571,590,821	0.63	₩53,635,640	0.63
sq	₩156,002,223	0.91	₩756,639,734	0.83	₩68,499,631	0.89
tsq	₩181,057,864	0.64	₩684,987,714	0.69	₩56,418,282	0.69
uy	₩613,290,211	0.93	₩578,919,118	0.87	₩176,154,652	0.89
ty	₩398,797,491	0.76	₩622,437,365	0.76	₩163,661,042	0.75

The forecasting methods minimising the total inventory costs in each equipment group are shown in bold.

Table 5-17 provides the mean ranks of direct forecasting methods for equipment group in terms of the total inventory costs and the mean rank for MAD (Table 5-13). Only the results of MAD were compared with the results of the total inventory costs, because the mean ranks for MAD were consistent with those for RMSE as shown in Table 5-13. As shown in Table 5-17, the mean ranks in terms of the total inventory costs presented consistent patterns with the mean ranks in terms of MAD. For Gun/RD, *um* minimised the mean ranks in terms of both the total inventory costs and MAD; for ME and GE/AC *tsm* minimised the mean ranks in terms of both the total inventory costs and MAD. As mentioned earlier, the different performance of the forecasting methods might originate from some of the differences of the data features in the equipment groups.

Table 5-17 Mean ranks of direct forecasting methods for equipment group in terms of inventory costs and MAD (Friedman's test, *p*-values are shown in the 3rd row)

	Gun/RD		ME		GE/AC	
	Total costs	Mean rank (MAD)	Total costs	Mean rank (MAD)	Total costs	Mean rank (MAD)
p-value	$p = 0.001$	$p = 0.002$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
um	**3.85**	**3.73**	5.45	5.02	5.32	5.07
tm	6.14	6.07	5.28	6.45	4.71	5.41
sm	5.91	5.20	5.77	5.15	5.94	5.41
tsm	5.45	5.41	**4.46**	**4.52**	**4.25**	**3.87**
uq	4.57	4.64	6.31	5.46	6.31	6.29
tq	6.27	6.55	5.21	6.40	4.60	5.75
sq	5.43	5.39	6.44	5.57	6.87	6.51
tsq	6.30	6.09	5.04	5.50	4.40	4.75
uy	5.68	5.93	5.29	4.99	5.35	5.47
ty	5.40	5.98	5.38	5.89	6.31	6.40

The mean rank of the forecasting method among the 10 direct forecasting methods in each equipment group in terms of the total inventory costs or MAD is presented in each cell; the forecasting methods minimising the mean rank in terms of the total inventory costs or MAD are shown in bold.

Table 5-18 Robust direct forecasting methods

	Absolute measure	Derivative measure
The most robust DF	*tsm*	*um* (total inventory costs); *tsm* (mean rank for total inventory costs)
Equipment group	Gun/RD: *um*;	ME & GE/AC: *tsm*

Table 5-18 presents the major findings of this section; that is, the robust direct forecasting methods in terms of absolute and derivative measures. There were some inconsistencies in the results of direct forecasting methods between absolute and derivative measures. In the years 2005 ~ 2007, while the forecast with monthly aggregated data adjusted for linear trend and additive seasonality (*tsm*) was found to be the most robust direct forecasting method in terms of MAD and RMSE, the forecast with monthly aggregated unadjusted data (*um*) was found to be the most robust direct forecasting method in terms of the total inventory costs. However, some consistent results were found in terms of the mean ranks; that is, *tsm* was the most robust direct forecasting method in terms of the mean ranks by MAD, RMSE, and the total inventory costs.

There were also some consistent results of direct forecasting methods for the equipment groups between absolute and derivative measures. *um* was the most robust direct forecasting method for Gun/RD, and *tsm* was the most robust direct forecasting method for the other two groups. The different forecasting performance in the different equipment groups might originate from the differences in data features identified in Subsection 4.3.4.

For the 300 items tested, the two direct forecasting methods, *um* and *tsm*, were found to minimise the total inventory costs and the mean rank in terms of the total inventory costs respectively (as shown in Table 5-15). As stated in Section 1.3, the main objective of this research is to compare alternative forecasting strategies (i.e. direct forecasting and hierarchical forecasting). Forecasting results using hierarchical forecasting methods are compared with these two direct forecasting methods in the next section.

5.5 Hierarchical Forecasting

In Section 2.4, arguments from the literature about the performance of hierarchical forecasting were presented. As mentioned in Section 4.6, several demand features for the Naval spare parts that could make hierarchical forecasting more accurate than direct forecasting were identified. This section develops a range of hierarchical forecasting methods. Then, the forecasts generated using hierarchical forecasting methods are compared with the forecasts generated using direct forecasting methods in terms of the absolute, relative, and derivative accuracy measures.

5.5.1 The development of hierarchical forecasting methods

Various proration methods were reviewed in Subsection 2.3.3 (Gross and Sohl, 1990, Fliedner and Lawrence, 1995, DeLurgio, 1998, Narasimhan et al., 1998, Fliedner, 1999, Fliedner, 2001, Widiarta et al., 2008b). Four proration methods capable of producing a long forecast horizon were employed. This is because the Naval procurement system requires a long forecast horizon as stated in Subsection 4.3.6.

As stated in Subsection 2.3.3, Gross and Sohl (1990) compared several top-down forecasting proration methods for predicting demand for three industrial galvanised steel product sales [e.g. equations (2-21), (2-22), (2-24), and (2-25)]. They found reasonably good performance with the two mean proportion methods [equations (2-21) and (2-22)]. This research employed these two kinds of top-down forecasting (TD) proration methods. TD1 produced the item level forecast by multiplying the group level direct forecast by the mean ratio of the item's demand to the group's demand as shown in equation (2-21). TD2 produced the item level forecast by multiplying the group level direct forecast by the ratio of the item's mean demand to the group's mean demand as shown in equation (2-22).

Combinatorial forecasting has been claimed to be a good forecasting method (DeLurgio, 1998, Kahn, 1998, Dekker et al., 2004, Hyndman et al., 2007). It could correct an irregularity of a time series consisting of two time series and improve the combinatorial forecasting accuracy (DeLurgio, 1998, Dekker et al., 2004). This research employed two combinatorial forecasting methods which were recommended by DeLurgio (1998). Simple combination (SC) is defined as in equation (2-27). While SC is the simplest combinatorial method, SC was claimed to be as good as other more sophisticated methods (DeLurgio, 1998). Weighted combination (WC) is defined as in equation (2-28) with weights defined as in equation (2-30). WC is a rational way to combine forecasting models, because WC sets higher weights into more accurate forecasting models. WC is based on the idea that forecasting accuracy and the sum of squared errors are inversely related.

Corresponding to the direct forecasts produced as shown in Subsection 5.4.1, the direct forecasts for the hierarchical forecasts were produced identically. 10 TD1 and 10 TD2 methods were generated with the 10 group level direct forecasts ($\widetilde{F}_{t+\tau}$). 100 SC and 100 WC methods were produced with the combinations of the 10 group level direct forecasts ($\widetilde{F}_{t+\tau}$) and the 10 item level direct forecasts ($\widetilde{f}_{i,t+\tau}$). In total 220 hierarchical forecasting methods were generated.

As with the direct forecasts, forecasting horizons for the hierarchical forecasts were also established as the period including PROLT and review period. The hierarchical forecasts were generated until November 2007; the performance of the hierarchical forecasts was measured until November 2007. After generating the forecasts using the adjusted data, the data were reverse adjusted so that the forecasts can be compared with actual demand.

5.5.2 Forecasting performance comparisons

In Subsection 3.5.2, it was stated that the logistical case of the South Korean Navy is a typical case as well as an extreme case. The two peak points might represent the extreme case which is typical of military logistics. This typical/extreme case could have a biasing effect upon the performance of the forecasting method. The high impact of the two peak points in 2002 and 2003 upon the forecasts in 2004 ~ 2007 was identified as shown in Subsection 5.4.2 and 5.4.3. In this subsection, the forecasting periods with the exception of 2004 were considered. The performance of the 220 hierarchical forecasting methods in 2005 ~ 2007 was compared with that of the most robust direct forecasting method (*tsm*) at item level in terms of MAD and RMSE.

Appendix D and Appendix E compare the performance of the 220 hierarchical forecasting methods. LN(ratio) denotes the natural log of the ratios. Appendix D presents LN(ratio) for MAD; Appendix E presents LN(ratio) for RMSE. In the forecasting period 2005 ~ 2007, the relative performance of the hierarchical forecasting methods was rather moderate. 36 hierarchical forecasting methods (16.4%) were superior to the most robust direct forecasting method (*tsm*) in terms of the LN(ratio) for MAD (as shown in Appendix D).

Although the forecasting results in terms of MAD and RMSE were similar, the top 20 hierarchical forecasting methods were different. In order to find a group of good forecasting methods in terms of both measures, the common top 21 hierarchical forecasting methods in terms of the both measures were selected. Table 5-19 presents the top 21 hierarchical forecasting methods in terms of the LN(ratio) for MAD in the forecasting period 2005 ~ 2007.

Table 5-19 Top 21 hierarchical forecasting methods (MAD)

	LN(ratio) for MAD			Rank		
	05 ~ 07	06 ~ 07	07	05 ~ 07	06 ~ 07	07
TD2tsm	-22.98	-17.62	22.77	6	9	42
SCtmum	-24.75	-7.50	43.33	5	24	93
SCtmsm	-21.87	-19.40	34.01	8	8	67
SCtmuq	-9.14	1.94	47.78	16	33	105
SCtsmum	-26.02	**-27.74**	2.79	3	**1**	8
SCtsmsm	-20.63	-24.98	15.64	9	3	28
SCtsmuq	-5.92	-1.82	13.09	21	29	22
SCtqum	**-32.10**	-12.71	33.56	**1**	13	62
SCtqsm	-25.25	-23.09	24.69	4	6	47
SCtquq	-15.07	-4.77	34.85	13	25	70
SCtqsq	-6.83	-8.95	37.86	20	18	81
SCtquy	-8.79	5.20	35.54	17	40	74
SCtsqum	-27.00	-20.55	7.75	2	7	10
SCtsqsm	-18.54	-23.69	15.28	10	5	25
SCtyum	-12.35	-23.79	26.86	14	4	52
WCtmum	-10.06	-3.32	11.38	15	27	19
WCtsmum	-15.46	-12.02	-1.57	12	14	4
WCtqum	-22.34	-15.87	-0.16	7	10	6
WCtqsm	-7.55	-15.22	-7.28	19	11	2
WCtsqum	-17.85	-14.54	-6.29	11	12	3
WCtyum	-7.61	-3.12	**-14.63**	18	28	**1**

The most robust forecasting methods are shown in bold; LN(ratio) = the sum of natural log relative error [$\ln(\text{error}_{HF}/\text{error}_{DF})$] over the 300 items.

In the forecasting period 2005 ~ 2007, 39 hierarchical forecasting methods (17.7%) were superior to *tsm* in terms of the LN(ratio) for RMSE as shown in Appendix E. Table 5-20 presents the top 21 hierarchical forecasting methods in terms of the LN(ratio) for RMSE in the forecasting period 2005 ~ 2007.

Table 5-20 Top 21 hierarchical forecasting methods (RMSE)

	LN(ratio) for RMSE			Rank		
	05 ~ 07	06 ~ 07	07	05 ~ 07	06 ~ 07	07
TD2tsm	-23.81	-19.41	22.77	8	9	42
SCtmum	-28.76	-11.85	43.33	3	16	93
SCtmsm	-25.35	-23.93	34.01	6	7	67
SCtmuq	-12.37	-0.22	47.78	14	32	105
SCtsmum	-27.68	**-27.80**	2.79	5	**1**	8
SCtsmsm	-22.71	-24.93	15.64	9	5	28
SCtsmuq	-8.04	-0.40	13.09	20	31	22
SCtqum	**-36.38**	-16.20	33.56	**1**	11	62
SCtqsm	-28.81	-26.74	24.69	2	3	47
SCtquq	-17.02	-5.16	34.85	13	26	70
SCtqsq	-9.25	-12.30	37.86	18	15	81
SCtquy	-12.13	5.61	35.54	17	41	74
SCtsqum	-28.66	-21.76	7.75	4	8	10
SCtsqsm	-20.30	-24.69	15.28	11	6	25
SCtyum	-12.33	-26.28	26.86	15	4	52
WCtmum	-12.32	-4.10	11.38	16	27	19
WCtsmum	-17.08	-11.78	-1.57	12	17	4
WCtqum	-24.34	-16.53	-0.16	7	10	6
WCtqsm	-8.99	-15.29	-7.28	19	12	2
WCtsqum	-20.80	-14.97	-6.29	10	13	3
WCtyum	-7.83	-0.87	**-14.63**	21	29	**1**

The most robust forecasting methods are shown in bold; LN(ratio) = the sum of natural log relative error [ln($error_{HF}$/$error_{DF}$)] over the 300 items.

As shown in Table 5-19 and Table 5-20, the performance of the top 21 hierarchical forecasting methods in 2005 ~ 2007 was more or less consistent with that in 2006 ~ 2007; that is, in terms of the LN(ratio) the top 21 hierarchical forecasting methods in the years 2005 ~ 2007 were also superior to *tsm* in the years 2006 ~ 2007 with the exception of *SCtmuq* and *SCtquy*. However, only five hierarchical forecasting methods of the top 21 hierarchical forecasting methods in the year 2007 were superior to *tsm*. An eleven month period in 2007 (i.e. January 2007 ~ November 2007) might be insufficient to compare the performance of the forecasting methods. Within the top 21, no TD1, 1 TD2, 14 (66.7%) SC and 6 WC methods are included as shown in Table 5-19 and Table 5-20. SC dominated the top 21.

There was some preference for employing data aggregation and data adjustment methods for the top 21 hierarchical forecasting methods. At group level, *m*, *q* and *y* were employed for 9, 10 and 2 hierarchical forecasting methods respectively as the data aggregation method; *t* and *ts* were employed for 13 and 8 hierarchical forecasting methods respectively as the data adjustment method. At item level, *m*, *q* and *y* were employed for 15, 4 and 1 hierarchical forecasting methods respectively as the data aggregation method; *u* and *s* were employed for 14 and 6 hierarchical forecasting methods respectively as the data adjustment method. At group level, *m* and *q* were the most frequently higher ranked data aggregation methods and *t* and *ts* were the most frequently higher ranked data adjustment methods for the top 21. At item level, *m* was the most frequently higher ranked data aggregation method and *u* was the most frequently higher ranked data adjustment method. As the combination of the data aggregation methods and the data adjustment methods, *tq* at group level and *um* at item level were the most frequently higher ranked methods. At group level *tq* was employed for 7 hierarchical forecasting methods; at item level *um* was utilised for 10 hierarchical forecasting methods.

The performance of the 220 hierarchical forecasting methods were compared with each other in Appendix F. Table 5-21 presents the mean ranks of the top 21 hierarchical forecasting methods between 2005 and 2007 in Appendix F. Friedman's test for forecasting errors was conducted. The number of treatments is 220 (i.e. the number of forecasting methods) and the number of blocks is 300 (i.e. the number of items). The top 21 forecasting methods in terms of the mean rank in Table 5-21 are different from the top 21 forecasting methods in terms of the LN(ratio) in Table 5-19 and Table 5-20. Three forecasting methods (i.e. *SCtqsq*, *SCtyum*, and *WCtyum*) included in the top 21 forecasts in terms of the LN(ratio) for MAD and RMSE are not included in the top 21 forecasts in terms of the mean rank for MAD and RMSE; three forecasting methods (i.e. *SCtmuy*, *SCtsmuy*, and *SCtsquy*) included in the top 21 forecasts in terms of the mean rank for MAD and RMSE are not included in the top 21 forecasts in terms of the LN(ratio) for MAD and RMSE. However, the test results in Table 5-21 were similar to the results in Table 5-19 and Table 5-20 in that the most robust forecasting method was consistently *SCtqum* from all the tests, and SC dominated the top 21. No TD1, 1 TD2, 15 (71.4%) SC and 5 WC methods included within the top 21 hierarchical forecasting methods in terms of the mean rank.

Table 5-21 Top 21 hierarchical forecasting methods (mean rank)
(Friedman's test, p-value < 0.001)

	MAD		RMSE	
	Mean rank	Rank	Mean rank	Rank
TD2tsm	84.90	17	85.26	19
SCtmum	79.34	4	79.21	7
SCtmsm	79.05	3	78.38	3
SCtmuq	84.92	18	83.93	17
SCtmuy	80.01	7	79.00	4
SCtsmum	78.70	2	79.07	5
SCtsmsm	81.30	10	81.03	9
SCtsmuq	85.77	19	85.81	20
SCtsmuy	84.05	15	81.67	11
SCtqum	77.54	**1**	77.37	**1**
SCtqsm	79.70	6	77.67	2
SCtquq	84.33	16	83.47	14
SCtquy	80.97	9	79.15	6
SCtsqum	80.95	8	81.61	10
SCtsqsm	83.85	14	83.74	16
SCtsquy	86.63	21	84.21	18
WCtmum	82.17	11	83.67	15
WCtsmum	82.52	12	82.93	13
WCtqum	79.64	5	79.73	8
WCtqsm	85.78	20	85.88	21
WCtsqum	83.39	13	82.83	12

Mean rank = mean rank of the forecasting method among the 220 hierarchical forecasting methods over the 300 items in terms of MAD or RMSE; Rank = the rank of the forecasting method among the 220 hierarchical forecasting methods based on the mean rank; the forecasting method minimising the mean rank are shown in bold.

A summary of direct forecasting methods at group and item levels used for the top 21 hierarchical forecasting methods is presented in Table 5-22. The frequencies of direct forecasting methods used for the top 21 hierarchical forecasting methods in terms of the LN(ratio) and the mean rank were similar.

Table 5-22 Direct forecasting methods used for the top 21 hierarchical forecasting methods

		LN(ratio)		Mean rank	
		Group	Item	Group	Item
Data Aggregation	*m*	9	15	11	13
	q	10	4	10	3
	y	2	1	.	4
Data Adjustment	*u*	.	14	.	15
	t	13	.	11	.
	s	.	6	.	5
	ts	8	.	10	.

The number of hierarchical forecasting methods in the top 21 which the direct forecasting method was used for either at group level or at item level is presented in each cell; the LN(ratio) from Table 5-19 and Table 5-20; the mean rank from Table 5-21.

At group level, *m* and *q* were the most frequently higher ranked data aggregation methods; and *t* and *ts* were the most frequently higher ranked data adjustment methods. At item level, *m* was the most frequently higher ranked data aggregation method; and *u* was the most frequently higher ranked data adjustment method. At item level, *m* was used for 15 (71.4%) and 13(61.9%) hierarchical forecasting methods in terms of the LN(ratio) and the mean rank respectively; *u* was used for 14 (66.7%) and 15 (71.4%) hierarchical forecasting methods in terms of the LN(ratio) and the mean rank respectively.

The combinations of the data aggregation methods and the data adjustment methods for the top 21 hierarchical forecasting methods were also investigated. At group level, *tq* was the most frequently higher ranked combination, that is, *tq* was used for 7 hierarchical forecasting methods in Table 5-19 and Table 5-20. In Table 5-21, at group level, while *tq* was used for 6 hierarchical forecasting methods, *tsm* and *tm* also used 6 and 5 hierarchical forecasting methods respectively. *tq* was not the only most frequently higher ranked combination at group level in Table 5-21. *um*, which was the most frequently higher ranked combination at item level in terms of MAD and RMSE, was also the most frequently higher ranked combination in terms of the mean rank as shown in Table 5-21. While *um* was used for 10 hierarchical forecasting methods at item level in Table 5-19 and Table 5-20, *um* was used for 8 hierarchical forecasting methods at item level in Table 5-21.

Table 5-23 compares the relative performance of the proration methods in the forecasting years 2005 ~ 2007. In order to calculate the mean LN(HF/*tsm*), the sum of the LN(HF/*tsm*) for each item using each forecasting method was divided by *n* (*n* = the number of item × the number of proration methods). For example, the sum of the LN(HF/*tsm*) for 300 items using TD1 was divided by 3,000 [3,000 = 300 (no. of the items) × 10 (no. of the TD1 proration methods)]; the sum of the LN(HF/*tsm*) for 300 items using SC was divided by 30,000 [300 (no. of the items) × 100 (no. of the SC proration methods)]. The standard deviation of the mean used the identical *n* which was used for calculating the mean LN(HF/*tsm*).

Table 5-23 Proration methods comparisons [LN(HF/*tsm*)]

	MAD		RMSE	
	Mean LN(HF/*tsm*)	Std	Mean LN(HF/*tsm*)	Std
TD1	0.22	0.09	0.20	0.08
TD2	0.09	0.07	0.08	0.07
SC	0.12	0.12	0.11	0.12
WC	0.11	0.08	0.10	0.08

Mean LN(HF/*tsm*) = the mean values of natural log relative error [$\ln(\text{error}_{HF}/\text{error}_{tsm})$] for MAD or RMSE per item for forecasting methods using the proration method over the 300 items; Std = standard deviation of the mean LN(HF/*tsm*).

SC dominated the top 21 in Table 5-19, Table 5-20 and Table 5-21. As shown in Table 5-23, TD2 presented the minimum mean LN(HF/*tsm*) and WC presented the second minimum mean LN(HF/*tsm*). The reason why TD2 presented the minimum mean LN(HF/*tsm*) than that of SC might originate from the higher standard deviation of the mean LN(HF/*tsm*) of SC. This implies that the performance of SC was highly variable.

Table 5-24 compares the performance of the proration methods in the 3 equipment groups in the forecasting period 2005 ~ 2007. *um* was compared with hierarchical forecasting methods in Gun/RD, and *tsm* was compared with hierarchical forecasting methods in ME and GE/AC. This is because *um* was the most robust direct forecasting method for Gun/RD, and *tsm* was the most robust direct forecasting method for ME and GE/AC as shown in Table 5-13. The LN(ratio) and the Std in Table 5-24 were calculated the same way as the mean LN(HF/*tsm*) and the Std in Table 5-23. For example, the sum of the natural log relative error [ln($error_{HF}/error_{DF}$)] for MAD or RMSE for 188 items for ME using SC was divided by 18,800 [188 (no. of items) × 100 (no. of proration methods)].

Table 5-24 Proration methods comparisons in the 3 equipment groups [LN(ratio)]

	MAD						RMSE					
	Gun/RD		ME		GE/AC		Gun/RD		ME		GE/AC	
	LN(ratio)	Std	LN(ratio)	Std	LN(ratio)	Std	LN(ratio)	Std	LN(ratio)	Std	LN(ratio)	Std
TD1	0.204	0.080	0.280	0.089	0.193	0.160	0.216	0.070	0.268	0.086	0.172	0.146
TD2	**0.097**	0.085	0.133	0.074	**0.133**	0.109	**0.113**	0.074	0.123	0.071	**0.104**	0.102
SC	0.243	0.189	0.138	0.123	0.135	0.114	0.263	0.186	0.134	0.123	0.121	0.110
WC	0.260	0.163	**0.115**	0.088	0.152	0.088	0.281	0.166	**0.113**	0.086	0.135	0.083

LN(ratio) = the mean values of natural log relative error [ln($error_{HF}/error_{DF}$)] for MAD or RMSE per item in the equipment group; Std = standard deviation of LN(ratio); the proration methods presenting the minimum LN (ratio) in each group are shown in bold.

TD2 presented the minimum LN(ratio) for Gun/RD and GE/AC, and WC presented the minimum LN(ratio) for ME as shown in Table 5-24. The standard deviations of the mean LN(ratio) of SC were the highest in Gun/RD and ME. Although SC dominated the top 21 in Table 5-19, Table 5-20 and Table 5-21, SC for each equipment group performed the least well. This might be caused by the highly variable performance of SC which can be identified by the high standard deviation as shown in Table 5-24.

It is necessary to examine the performance of the hierarchical forecasting methods for the each group. *um* was compared with the hierarchical forecasting methods for Gun/RD, and *tsm* was compared with the hierarchical forecasting methods for ME and GE/AC. Among the 220 hierarchical forecasting methods, 18 (8.2%) and 14 (6.4%) hierarchical forecasting methods were superior to *um* for Gun/RD in terms of LN(HF/*um*) for MAD and RMSE respectively; 28 (12.7%) hierarchical forecasting methods were superior to *tsm* for ME in terms of both LN(HF/*tsm*) for MAD and RMSE; and 18 (8.2%) and 20 (9.1%) hierarchical forecasting methods were superior to *tsm* for GE/AC in terms of the LN(HF/*tsm*) for MAD and RMSE respectively. Hierarchical forecasting provided much superior performance for ME than that for Gun/RD and GE/AC.

Table 5-25, Table 5-26, and Table 5-27 present the top 10 hierarchical forecasting methods for Gun/RD, ME, and GE/AC in terms of either LN(HF/DF) for MAD or LN(HF/DF) for RMSE respectively. SC dominated all the top 10 hierarchical forecasting methods. 7 (58.3%) SC methods are included in the 12 forecasting methods for Gun/RD (Table 5-25); 9 (81.8%) SC methods are included in the 11 forecasting methods for ME (Table 5-26); and 9 (81.8%) SC methods are included in the 11 forecasting methods for GE/AC (Table 5-27). The most robust forecasting method (*SCtqum*) for the 300 items tested is included in all the top 10 hierarchical forecasting methods for the 3 equipment groups.

Table 5-25 Top 10 hierarchical forecasting methods for Gun/RD in terms of either the natural log relative error for MAD or RMSE

	MAD		RMSE	
	LN(HF/*um*)	Rank	LN(HF/*um*)	Rank
TD2um	-1.5	12	-0.7	9
TD2tsm	-2.5	6	-0.2	13
SCtmum	-5.9	1	-4.8	1
SCtmsm	-4.0	3	-2.1	4
SCtsmum	-3.9	4	-2.6	3
SCtqum	-4.3	2	-3.8	2
SCtqsm	-2.4	7	-0.5	11
SCtsqum	-2.1	10	-1.0	8
SCtyum	-2.3	8	-1.6	6
WCtmum	-3.1	5	-2.0	5
WCtsmum	-1.8	11	-0.7	10
WCtqum	-2.3	9	-1.5	7

LN(HF/*um*) = the sum of natural log relative error [ln($error_{HF}/error_{um}$)] for MAD or RMSE of each forecasting method over the items in Gun/RD; Rank = the rank of the forecasting method in the 220 forecasting methods in terms of LN(HF/*um*).

Table 5-26 Top 10 hierarchical forecasting methods for ME in terms of either the natural log relative error for MAD or RMSE

	MAD		RMSE	
	LN(HF/*tsm*)	Rank	LN(HF/*tsm*)	Rank
SCtmuy	-7.2	8	-8.1	10
SCtsmum	-11.5	3	-11.2	4
SCtsmsm	-7.0	10	-8.6	9
SCtsmuy	-7.1	9	-6.8	13
SCtqum	-10.9	4	-12.8	2
SCtqsm	-6.4	13	-9.7	6
SCtquy	-12.6	1	-13.1	1
SCtsqum	-11.6	2	-11.4	3
SCtsquy	-7.3	7	-8.8	8
WCtqum	-10.1	6	-9.7	7
WCtsqum	-10.6	5	-10.8	5

LN(HF/*tsm*) = the sum of natural log relative error [ln($error_{HF}/error_{tsm}$)] for MAD or RMSE of each forecasting method over the items in ME; Rank = the rank of the forecasting method in the 220 forecasting methods in terms of LN(HF/*tsm*).

Table 5-27 Top 10 hierarchical forecasting methods for GE/AC in terms of either the natural log relative error for MAD or RMSE

	MAD		RMSE	
	LN(HF/*tsm*)	Rank	LN(HF/*tsm*)	Rank
TD1tsm	-3.0	8	-3.8	7
TD2tsm	-3.3	6	-4.3	5
SCtmum	-3.4	5	-5.1	4
SCtmsm	-5.5	3	-5.9	3
SCtsmsm	-2.2	10	-2.4	11
SCtqum	-6.6	1	-7.9	1
SCtqsm	-6.2	2	-6.6	2
SCtquq	-2.7	9	-3.1	10
SCtquy	-1.7	11	-3.7	8
SCtsqum	-3.0	7	-4.2	6
SCtsqsm	-3.5	4	-3.7	9

LN(HF/*tsm*) = the sum of natural log relative error [$\ln(error_{HF}/error_{tsm})$] for MAD or RMSE of each forecasting method over the items in GE/AC; Rank = the rank of the forecasting method in the 220 forecasting methods in terms of LN(HF/*tsm*).

Table 5-28 presents the major findings of this subsection; that is, the performance of hierarchical forecasting methods in terms of absolute and relative measures. In this subsection, the 220 hierarchical forecasting methods were compared with the most robust direct forecasting method (*tsm*) in terms of the natural log relative error. 36 (16.4%) and 39 (17.7%) hierarchical forecasting methods were superior to *tsm* in terms of the LN(ratio) for MAD and RMSE in the years 2005 ~ 2007 respectively.

The most robust forecasting method for the 300 items was found to be *SCtqum* in terms of the LN(ratio) and the mean rank in the years 2005 ~ 2007. Combinational forecasting (especially, simple combination) dominated the top 21 hierarchical forecasting methods ranked by the LN(ratio). The most frequently higher ranked data aggregation and data adjustment methods for the top 21 hierarchical forecasting methods were analysed in this subsection.

Table 5-28 The performance of hierarchical forecasting methods in terms of absolute and relative measures

	Performance
Superior to the most robust DF	16.4% & 17.7% of HFs to *tsm* by LN(ratio) for MAD & RMSE
The most robust forecasting method	*SCtqum*
Proration method	Top-21 by LN(ratio): 14 SCs, 6 WCs, 0 TD1, & 1 TD2 Top-21 by mean rank in MAD/RMSE: 15 SCs, 5 WCs, 0 TD1, & 1 TD2
The most frequently higher ranked DF for Top-21	Group level: m^1 & q^1; t^2& ts^2 Group level combination: *tq* Item level: m^1 & u^2 Item level combination: *um*

DF = direct forecasting; HF = hierarchical forecasting; SC = simple combination; WC = weighted combination; TD = top-down forecasting; *1* = data aggregation method; and *2* = data adjustment method; LN(ratio) = natural log relative error [$\ln(\text{error}_{HF}/\text{error}_{DF})$].

5.5.3 Simulation

In the above subsection, the 220 hierarchical forecasting methods were compared with the most robust direct forecasting method (*tsm*) identified by MAD and RMSE in terms of the absolute and relative measures. However, as stated in Section 2.6, these measures do not present the practical impact of a forecasting method upon the inventory system. In this section, direct forecasting methods are compared with the 220 hierarchical forecasting methods in terms of the total inventory costs and the inventory fill rate derived from simulations.

The total inventory costs of the 220 hierarchical forecasting methods were compared with *um*; the relative total inventory costs of the hierarchical forecasting methods were compared with *tsm* using the LN(ratio). This is because *um* was the most robust direct forecasting method in terms of the total inventory costs; *tsm* was the most robust direct forecasting method in terms of the mean rank for the total inventory costs as shown in Table 5-15. The entire simulation results for the 300 items between 2005 and 2007 are presented in Appendix G. 35 (15.9%) forecasting methods of the 220 hierarchical forecasting methods were superior to *um* in terms of the total inventory costs as shown.

Table 5-29 Top 20 hierarchical forecasting methods

	MAD		RMSE		Total inventory costs			Fill rate
	Mean rank	Rank	Mean rank	Rank	Costs	HF-*um*	Rank	
SCtmum	79.34	4	79.21	7	₩726,002,802	-₩52,939,451	7	0.79
SCtmsm	79.05	3	78.38	3	₩730,726,074	-₩48,216,179	10	0.79
SCtmuq	84.92	18	83.93	17	₩740,191,882	-₩38,750,371	12	0.78
SCtsmum	78.70	2	79.07	5	₩730,245,647	-₩48,696,606	9	0.83
SCtsmsm	81.30	10	81.03	9	₩748,440,413	-₩30,501,840	16	0.83
SCtsmuq	85.77	19	85.81	20	₩748,193,308	-₩30,748,944	15	0.82
SCtqum	77.54	1	77.37	1	₩693,601,747	-₩85,340,505	1	0.79
SCtqsm	79.70	6	77.67	2	₩702,219,843	-₩76,722,409	3	0.79
SCtquq	84.33	16	83.47	14	₩706,669,261	-₩72,272,992	4	0.78
SCtyum	89.50	30	92.56	45	₩702,216,842	-₩76,725,411	2	0.83
SCtysm	92.13	43	95.40	55	₩715,748,865	-₩63,193,388	6	0.83
SCtyuq	95.07	53	98.77	59	₩714,672,940	-₩64,269,313	5	0.82
WCumtsm	89.00	25	88.05	24	₩758,279,736	-₩20,662,517	18	0.79
WCumtq	123.33	166	121.46	150	₩761,152,212	-₩17,790,041	20	0.67
WCtmum	82.17	11	83.67	15	₩756,997,140	-₩21,945,113	17	0.83
WCtsmum	82.52	12	82.93	13	₩747,740,810	-₩31,201,442	14	0.85
WCtqum	79.64	5	79.73	8	₩728,256,615	-₩50,685,638	8	0.83
WCtquq	90.18	33	91.14	38	₩759,015,207	-₩19,927,046	19	0.82
WCuytsm	89.11	28	90.40	35	₩745,773,185	-₩33,169,068	13	0.78
WCtyum	93.90	47	95.21	53	₩739,345,856	-₩39,596,397	11	0.81

Mean rank = mean rank of the forecasting method among the 220 hierarchical forecasting methods over the 300 items in terms of MAD or RMSE; Rank = the rank of the forecasting method among the 220 hierarchical forecasting methods; HF – *um* = the total inventory costs of each hierarchical forecasting method deducted by the total inventory costs of *um*; Fill rate = the mean inventory fill rate using the hierarchical forecasting method.

Of the 35 forecasting methods, the top 20 ranking forecasting methods in terms of the total inventory costs are presented in Table 5-29. Although 3 TD2 methods (i.e. *TD2um*, *TD2tsm* and *TD2uy*) were included in the above 35 hierarchical forecasting methods as shown in Appendix G, no TD1, no TD2, 12 SC and 8 WC methods are included in the top 20 hierarchical forecasting methods as shown in Table 5-29. This was consistent with the top 21 hierarchical forecasting methods in the previous subsection. Recalling the results in the previous subsection, no TD1, 1 TD2, 14 SC and 6 WC methods were included in the top 21 hierarchical forecasting methods in terms of the LN(ratio); no TD1, 1 TD2, 15 SC and 5 WC methods were included in the top 21 hierarchical forecasting methods in terms of the mean rank based on MAD and RMSE.

Table 5-29 compares the performance of the top 20 hierarchical forecasting methods from simulation with the performance of the forecasting methods in terms of MAD and RMSE which were presented in the previous subsection. Although, the ranks in MAD and RMSE and the total inventory costs were more or less inconsistent, there was a tendency that a good forecasting method in terms of MAD and RMSE was also a good forecasting method in terms of the total inventory costs. *SCtqum,* which held the rank of 1st by MAD and RMSE, held the rank of 1st by the total inventory costs. *SCtqsm*, which also held high ranks (i.e. the rank 6th by MAD and the rank 2nd by RMSE), held the rank of 3rd by the total inventory costs. With the exception of *WCumtq*, the top 20 forecasting methods in terms of the total inventory costs are included in the top 53 forecasting methods in terms of MAD and the top 59 forecasting methods in terms of RMSE. Although the exceptional case, *WCumtq*, is included in the top 20 forecasting method, this forecasting method was disqualified in terms of the inventory fill rate of less than 70% (i.e. 67%).

Direct forecasting methods at group and item levels used for the top 21 hierarchical forecasting methods in terms of the LN(ratio) and the mean rank and the top 20 hierarchical forecasting methods in terms of the total inventory costs are presented in Table 5-30. At group level, *m* as the data aggregation method and *t* as the data adjustment method were the most frequently higher ranked methods in terms of the total inventory costs as well as the LN(ratio) and the mean rank. At item level, *m* was the most frequently higher ranked data aggregation method and *u* was the most frequently higher ranked data adjustment method. At item level, *m* was used for 15 (71.4%), 13 (61.9%), and 14 (70.0%) hierarchical forecasting methods in terms of the LN(ratio), the mean rank, and the total inventory costs respectively; *u* was used for 14 (66.7%), 15 (71.4%), and 13 (65.0%) hierarchical forecasting methods in terms of the LN(ratio), the mean rank, and the total inventory costs respectively.

Table 5-30 Direct forecasting methods used for the top 21 or the top 20 hierarchical forecasting methods

		LN(ratio)		Mean rank		Total inventory costs	
		Group	Item	Group	Item	Group	Item
Data Aggregation	*m*	9	15	11	13	10	14
	q	10	4	10	3	5	6
	y	2	1	.	4	5	.
Data Adjustment	*u*	.	14	.	15	3	13
	t	13	.	11	.	13	1
	s	.	6	.	5	.	4
	ts	8	.	10	.	4	2

The number of hierarchical forecasting methods in the top 21 or the top 20 which the direct forecasting method was used for either at group level or at item level is presented in each cell; the LN(ratio) from Table 5-19 and Table 5-20; the mean rank from Table 5-21; the total inventory costs from Table 5-29.

As for the combination of the data aggregation methods and the data adjustment methods, recalling the results in the previous subsection, *um* was employed for 10 (47.6%) and 8 (38.1%) hierarchical forecasting methods at item level in terms of the LN(ratio) (Table 5-19 and Table 5-20) and the mean rank (Table 5-21) respectively. *um* was still the most frequently higher ranked method at item level in terms of the total inventory costs as *um* was employed for 8 (40%) hierarchical forecasting methods of the top 20 hierarchical forecasting methods as shown in Table 5-29.

Appendix G compares the performance of direct and hierarchical forecasting methods in terms of the total inventory costs and the relative total inventory costs. While 35 forecasting methods of the 220 hierarchical forecasting methods were superior to *um* in terms of the total inventory costs, 19 forecasting methods of the 220 hierarchical forecasting methods were superior to *tsm* in terms of the LN(HF/*tsm*) for the total inventory costs [equation (5-6)]. In order to establish internal validity the 220 hierarchical forecasting methods were assessed with both of the measures.

Table 5-31 presents the 33 hierarchical forecasting methods which include the above top 20 hierarchical forecasting methods in Table 5-29 as well as the 19 superior hierarchical forecasting methods in terms of the LN (HF/*tsm*) for the total inventory costs. The 1st ranked forecasting method in terms of the total inventory costs, *SCtqum* was found to be the 2nd ranked forecasting method in terms of the LN(HF/*tsm*) for the total inventory costs. On the other hand, *TD2tsm*, which ranked as 27th in terms of the total inventory costs, was observed as the 1st ranked hierarchical forecasting method in terms of LN (HF/*tsm*) for the total inventory costs.

Table 5-31 The 33 hierarchical forecasting methods in simulation

	MAD		RMSE		Total inventory costs				
	Mean rank	Rank	Mean rank	Rank	Total inventory costs	HF-*um*	Rank	LN (HF/*tsm*)	Rank
TD2tm	118.60	137	116.04	123	₩854,063,515	₩75,121,262	109	-0.98	19
TD2tsm	84.90	17	85.26	19	₩770,753,404	-₩8,188,848	27	-25.38	1
TD2tq	122.44	157	120.02	144	₩799,522,960	₩20,580,707	65	-3.67	16
SCtmum	79.34	4	79.21	7	₩726,002,802	-₩52,939,451	7	-18.42	3
SCtmsm	79.05	3	78.38	3	₩730,726,074	-₩48,216,179	10	-11.79	8
SCtmtsm	99.25	60	98.78	60	₩803,679,850	₩24,737,597	72	-13.54	6
SCtmuq	84.92	18	83.93	17	₩740,191,882	-₩38,750,371	12	-2.66	17
SCtsmum	78.70	2	79.07	5	₩730,245,647	-₩48,696,606	9	1.29	21
SCtsmsm	81.30	10	81.03	9	₩748,440,413	-₩30,501,840	16	9.99	33
SCtsmtsm	90.71	35	89.68	32	₩780,007,802	₩1,065,549	37	-8.28	11
SCtsmuq	85.77	19	85.81	20	₩748,193,308	-₩30,748,944	15	14.62	41
SCtqum	77.54	1	77.37	1	₩693,601,747	-₩85,340,505	1	-24.64	2
SCtqsm	79.70	6	77.67	2	₩702,219,843	-₩76,722,409	3	-16.08	5
SCtqtsm	101.74	66	99.43	61	₩773,325,925	-₩5,616,328	30	-17.04	4
SCtquq	84.33	16	83.47	14	₩706,669,261	-₩72,272,992	4	-9.04	10
SCtquy	80.97	9	79.15	6	₩991,497,949	₩212,555,696	158	-1.14	18
SCtsqum	80.95	8	81.61	10	₩833,085,383	₩54,143,131	92	-3.90	15
SCtsqtsm	97.88	59	97.24	58	₩819,406,914	₩40,464,661	82	-7.31	12
SCtyum	89.50	30	92.56	45	₩702,216,842	-₩76,725,411	2	13.53	39
SCtysm	92.13	43	95.40	55	₩715,748,865	-₩63,193,388	6	19.80	58
SCtyuq	95.07	53	98.77	59	₩714,672,940	-₩64,269,313	5	22.51	65
WCumtsm	89.00	25	88.05	24	₩758,279,736	-₩20,662,517	18	10.45	34
WCumtq	123.33	166	121.46	150	₩761,152,212	-₩17,790,041	20	26.52	78
WCtmum	82.17	11	83.67	15	₩756,997,140	-₩21,945,113	17	14.70	42
WCtmtsm	93.85	46	92.41	44	₩796,912,497	₩17,970,244	60	-6.91	13
WCtsmum	82.52	12	82.93	13	₩747,740,810	-₩31,201,442	14	15.07	45
WCtsmtsm	91.71	41	89.98	34	₩787,884,582	₩8,942,330	51	-6.33	14
WCtqum	79.64	5	79.73	8	₩728,256,615	-₩50,685,638	8	5.29	27
WCtqtsm	94.68	51	93.20	46	₩782,633,097	₩3,690,845	43	-12.68	7
WCtquq	90.18	33	91.14	38	₩759,015,207	-₩19,927,046	19	24.45	73
WCtsqtsm	95.13	54	93.79	48	₩836,978,887	₩58,036,634	96	-9.39	9
WCuytsm	89.11	28	90.40	35	₩745,773,185	-₩33,169,068	13	11.02	35
WCtyum	93.90	47	95.21	53	₩739,345,856	-₩39,596,397	11	5.38	28

Mean rank = mean rank for the forecasting method in terms of MAD or RMSE over the 300 items; Rank = the rank of the forecasting method among the 220 hierarchical forecasting methods; HF – *um* = the total inventory costs of each hierarchical forecasting method deducted by the total inventory costs of *um*; LN(HF/*tsm*) = the sum of natural log relative error [$\ln(\text{error}_{HF}/\text{error}_{tsm})$] for the total inventory costs of each forecasting method over the 300 items.

Friedman's test for the LN(HF/*tsm*) for the total inventory costs of the 33 hierarchical forecasting methods was conducted. Table 5-32 compares the total inventory costs, the LN(HF/*tsm*) for the total inventory costs and the mean rank according to the LN(HF/*tsm*) for the total inventory costs of the 33 hierarchical forecasting methods.

Table 5-32 Mean rank of the 33 hierarchical forecasting methods

	Total inventory costs		LN(HF/*tsm*)		Friedman's test (p<0.001)	
	Sum	Rank	Sum	Rank	Mean rank	Rank
TD2tm	₩854,063,515	109	-0.98	19	17.27	20
TD2tsm	₩770,753,404	27	-25.38	1	15.56	7
TD2tq	₩799,522,960	65	-3.67	16	17.48	22
SCtmum	₩726,002,802	7	-18.42	3	14.47	4
SCtmsm	₩730,726,074	10	-11.79	8	15.21	6
SCtmtsm	₩803,679,850	72	-13.54	6	16.58	18
SCtmuq	₩740,191,882	12	-2.66	17	16.07	11
SCtsmum	₩730,245,647	9	1.29	21	16.50	16
SCtsmsm	₩748,440,413	16	9.99	33	17.55	23
SCtsmtsm	₩780,007,802	37	-8.28	11	16.18	14
SCtsmuq	₩748,193,308	15	14.62	41	17.79	25
SCtqum	₩693,601,747	1	-24.64	2	13.49	1
SCtqsm	₩702,219,843	3	-16.08	5	14.32	3
SCtqtsm	₩773,325,925	30	-17.04	4	16.14	13
SCtquq	₩706,669,261	4	-9.04	10	14.99	5
SCtquy	₩991,497,949	158	-1.14	18	14.13	2
SCtsqum	₩833,085,383	92	-3.90	15	15.80	9
SCtsqtsm	₩819,406,914	82	-7.31	12	17.19	19
SCtyum	₩702,216,842	2	13.53	39	18.25	29
SCtysm	₩715,748,865	6	19.80	58	19.12	31
SCtyuq	₩714,672,940	5	22.51	65	19.30	32
WCumtsm	₩758,279,736	18	10.45	34	17.82	26
WCumtq	₩761,152,212	20	26.52	78	19.32	33
WCtmum	₩756,997,140	17	14.70	42	17.34	21
WCtmtsm	₩796,912,497	60	-6.91	13	16.04	10
WCtsmum	₩747,740,810	14	15.07	45	18.01	27
WCtsmtsm	₩787,884,582	51	-6.33	14	16.25	15
WCtqum	₩728,256,615	8	5.29	27	16.52	17
WCtqtsm	₩782,633,097	43	-12.68	7	15.73	8
WCtquq	₩759,015,207	19	24.45	73	18.36	30
WCtsqtsm	₩836,978,887	96	-9.39	9	16.09	12
WCuytsm	₩745,773,185	13	11.02	35	17.76	24
WCtyum	₩739,345,856	11	5.38	28	18.04	28

Mean rank = mean rank for the forecasting method in terms of the LN(HF/*tsm*) for the total inventory costs over the 300 items.

The robustness of *SCtqum* in terms of the total inventory costs was confirmed by the mean rank. The 1st ranked *TD2tsm* in terms of the LN(HF/*tsm*) for the total inventory costs was merely ranked as 7th by the mean rank.

Table 5-33 compares the performance of the proration methods for the 300 items in terms of derivative and relative measures in the forecasting years 2005 ~ 2007. The mean values (i.e. mean inventory costs, mean stock, mean back order, and mean fill rate) and the standard deviation of mean used the identical *n* used for calculating the mean LN(HF/*tsm*) in Table 5-23.

As shown in Table 5-23 the LN(HF/*tsm*) for MAD and the LN(HF/*tsm*) for RMSE were found to be similar to each other. Thus, the only result from MAD is provided. In terms of both the total inventory costs and the mean LN(HF/*tsm*) for MAD, TD2 presented the minimum mean total inventory costs and the minimum mean LN(HF/*tsm*).

Table 5-33 Proration methods comparisons (simulation)

	Mean inventory costs		Mean	Mean	Mean	Mean LN(HF/*tsm*)
	Mean	Std	stock	backorder	fill rate	for MAD
TD1	₩4,549,086	₩695,251	197	26	0.78	0.22
TD2	₩2,877,407	₩424,018	194	24	0.76	0.09
SC	₩3,117,380	₩597,222	189	20	0.78	0.12
WC	₩3,025,997	₩599,491	182	22	0.78	0.11

Mean = the mean inventory costs per item for forecasting methods using the proration method over the 300 items; Std = the standard deviation of mean; mean stock (or backorder) = the mean stock (or back order) per item; mean LN(HF/*tsm*) for MAD = the mean values of natural log relative error [ln($error_{HF}$/$error_{tsm}$)] for MAD per item for forecasting methods using the proration method over the 300 items.

TD2 (which presented the minimum mean total inventory costs) was characterised as having the minimum standard deviation for the total inventory costs as shown in Table 5-33. Although there was no TD2 included within the top 20 hierarchical forecasting methods in Table 5-29, this minimum standard deviation of the mean inventory costs for TD2 might make TD2 present the minimum mean total inventory costs in Table 5-33. The high variability in the performance of SC was identified by the highest standard deviation of the mean LN(HF/*tsm*) for MAD and RMSE of SC as shown in Table 5-23.

The higher standard deviation of the mean total inventory costs of SC than that of TD2 was also identified as shown in Table 5-33.

It should be noted that SC produced the minimum mean backorder with the second lowest mean stock as shown in Table 5-33. Recalling equation (5-6) weighing twice the inventory carrying costs on the inventory stock-out costs, the equation can account for the reason why SC was observed to dominate the top 20 hierarchical forecasting methods in terms of the total inventory costs as shown in Table 5-29. In addition, there was no significant difference in the mean fill rate across the proration methods.

Table 5-34 Proration methods comparisons in the 3 equipment groups (simulation)

		Gun/RD	ME	GE/AC
TD1	Mean inventory costs	₩3,589,861	₩5,772,497	₩1,781,231
	Mean fill rate	0.75	0.76	0.85
	Mean LN(HF/DF) MAD	0.204	0.280	0.193
TD2	Mean inventory costs	₩3,469,953	₩3,483,099	₩814,288
	Mean fill rate	0.76	0.74	0.80
	Mean LN(HF/DF) MAD	0.097	0.133	0.133
SC	Mean inventory costs	₩4,463,679	₩3,543,689	₩1,062,390
	Mean fill rate	0.77	0.77	0.80
	Mean LN(HF/DF) MAD	0.243	0.138	0.135
WC	Mean inventory costs	₩4,955,543	₩3,238,745	₩1,183,738
	Mean fill rate	0.77	0.77	0.80
	Mean LN(HF/DF) MAD	0.260	0.115	0.152

Mean inventory costs = the mean inventory costs per item of the forecasting methods using the proration method in the equipment group; Std = the standard deviation of the mean inventory costs; mean LN(HF/DF) = the mean values of natural log relative error [$\ln(error_{HF}/error_{DF})$] for MAD per item in the equipment group.

Table 5-34 compares the performance of the proration methods in the 3 equipment groups in terms of the mean inventory costs and the LN(HF/DF) for MAD in the forecasting years 2005 ~ 2007. *um* was compared with the hierarchical forecasting methods in Gun/RD and *tsm* was compared with the hierarchical forecasting methods in ME and GE/AC. For Gun/RD and GE/AC, TD2 presented the minimum mean inventory costs and the minimum LN(HF/DF) for MAD; For ME, WC presented the minimum mean inventory costs and the minimum LN(HF/DF) for MAD. Most of the proration methods produce similar mean inventory fill rate. TD1 in GE/AC produces the highest mean inventory fill rate, at 85%.

The 220 hierarchical forecasting methods were compared with the most robust direct forecasting methods for the groups in terms of the relative measure in the previous subsection. The 220 hierarchical forecasting methods were compared with the best direct forecasting method for spare parts for each equipment group in terms of the derivative measure (i.e. the total inventory costs) in this subsection. *um* was compared with the hierarchical forecasting methods in Gun/RD and *tsm* was compared with the hierarchical forecasting methods in ME and GE/AC. This is because these direct forecasting methods were found to be the most robust direct forecasting methods for spare parts for the equipment groups in terms of both the relative and derivative measures.

In terms of the total inventory costs, among the 220 hierarchical forecasting methods, 6 (2.7%) hierarchical forecasting methods were superior to *um* for Gun/RD; 39 (17.7%) hierarchical forecasting methods were superior to *tsm* for ME; and 65 (29.5%) hierarchical forecasting methods were superior to *tsm* for GE/AC. Table 5-35, Table 5-36 and Table 5-37 present the top 10 hierarchical forecasting methods for spare parts for each equipment group in terms of either the sum of the LN(ratio) for MAD or RMSE, or the total inventory costs.

Table 5-35 Top 10 hierarchical forecasting methods for Gun/RD in terms of either the natural log relative error for MAD or RMSE, or the total inventory costs

	MAD		RMSE		Simulation					
	LN(ratio)	Rank	LN(ratio)	Rank	Total inventory costs	HF-*um*	Rank	Mean Stock	Mean Backorder	Fill rate
TD2um	-1.5	12	-0.7	9	₩119,228,678	₩4,187,708	13	1134.18	21.36	0.91
TD2tsm	-2.5	6	-0.2	13	₩153,527,634	₩38,486,664	104	817.39	294.10	0.74
SCtmum	-5.9	1	-4.8	1	₩117,987,344	₩2,946,374	10	704.95	216.40	0.78
SCtmsm	-4.0	3	-2.1	4	₩121,397,922	₩6,356,952	17	844.60	143.34	0.83
SCtsmum	-3.9	4	-2.6	3	₩122,417,637	₩7,376,667	18	861.86	135.12	0.85
SCtqum	-4.3	2	-3.8	2	₩122,602,086	₩7,561,116	19	703.29	218.18	0.78
SCtqsm	-2.4	7	-0.5	11	₩126,163,277	₩11,122,307	24	848.46	143.65	0.83
SCtsqum	-2.1	10	-1.0	8	₩128,037,489	₩12,996,519	32	811.32	180.83	0.83
SCuyum	0.5	21	0.8	16	₩114,878,374	-₩162,596	6	755.07	203.81	0.81
SCuysm	1.2	26	2.4	26	₩113,686,365	-₩1,354,605	3	885.71	129.92	0.86
SCuyuq	4.5	54	5.4	55	₩117,386,011	₩2,345,041	7	780.76	201.28	0.82
SCtyum	-2.3	8	-1.6	6	₩113,472,778	-₩1,568,192	2	690.19	201.09	0.82
SCtysm	-0.2	17	1.2	18	₩114,648,085	-₩392,885	4	837.83	131.94	0.86
SCtyuq	3.0	37	3.8	36	₩117,425,104	₩2,384,134	8	723.38	205.40	0.80
WCumum	-0.4	16	-0.2	14	₩114,752,663	-₩288,307	5	1054.24	50.01	0.90
WCtmum	-3.1	5	-2.0	5	₩120,766,521	₩5,725,551	16	801.01	149.97	0.83
WCtsmum	-1.8	11	-0.7	10	₩122,840,265	₩7,799,295	21	883.03	116.30	0.86
WCtqum	-2.3	9	-1.5	7	₩122,642,884	₩7,601,914	20	769.27	170.42	0.81
WCuyum	1.7	28	2.4	27	₩117,801,524	₩2,760,554	9	743.21	156.80	0.82
WCuysm	3.0	38	3.7	35	₩111,534,862	-₩3,506,108	1	925.91	47.82	0.88

LN(ratio) = the sum of natural log relative error [$\ln(\text{error}_{HF}/\text{error}_{um})$] for MAD or RMSE of each forecasting method over the items in Gun/RD; HF-*um* = the total inventory costs of each hierarchical forecasting method deducted by the total inventory costs of *um*; Rank = the rank of the forecasting method in the 220 forecasting methods.

Consistent with the above results, SC dominated the top 10 hierarchical forecasting methods for Gun/RD and ME in terms of the total inventory costs. 7 and 8 SC methods are included in the top 10 forecasting methods for Gun/RD and ME respectively in terms of the total inventory costs. However, for GE/AC only 4 SC methods are included in the top 10 forecasting methods in terms of the total inventory costs.

Table 5-36 Top 10 hierarchical forecasting methods for ME in terms of either the natural log relative error for MAD or RMSE, or the total inventory costs

	MAD		RMSE		Simulation					
	LN(ratio)	Rank	LN(ratio)	Rank	Total inventory costs	HF-*tsm*	Rank	Mean Stock	Mean Backorder	Fill rate
SCtmuy	-7.2	8	-8.1	10	₩528,430,080	-₩41,245,186	8	50966.71	3931.64	0.79
SCtsmum	-11.5	3	-11.2	4	₩550,407,790	-₩19,267,476	17	46129.31	2305.95	0.82
SCtsmsm	-7.0	10	-8.6	9	₩556,323,892	-₩13,351,374	23	42651.41	2934.44	0.80
SCtsmuy	-7.1	9	-6.8	13	₩526,626,765	-₩43,048,501	7	45583.40	2478.43	0.82
SCtqum	-10.9	4	-12.8	2	₩517,805,384	-₩51,869,882	4	40917.39	3519.23	0.78
SCtqsm	-6.4	13	-9.7	6	₩521,395,839	-₩48,279,427	5	36803.17	3915.59	0.77
SCtquq	-2.3	20	-3.6	19	₩524,765,817	-₩44,909,449	6	41844.69	3699.22	0.77
SCtquy	-12.6	1	-13.1	1	₩491,733,078	-₩77,942,188	1	39963.45	4169.75	0.79
SCtsqum	-11.6	2	-11.4	3	₩647,552,177	₩77,876,911	144	41640.18	2798.17	0.80
SCtsquy	-7.3	7	-8.8	8	₩621,769,254	₩52,093,988	124	40902.33	3159.02	0.81
SCtyum	-0.4	24	0.7	31	₩530,764,346	-₩38,910,920	9	48839.60	2182.89	0.82
SCtyuy	2.4	38	3.7	44	₩510,603,793	-₩59,071,473	2	48437.87	2378.10	0.83
WCtqum	-10.1	6	-9.7	7	₩546,635,272	-₩23,039,994	14	44412.86	3077.09	0.81
WCtquy	-0.1	27	2.0	38	₩537,217,736	-₩32,457,530	10	52781.37	2143.35	0.84
WCtsqum	-10.6	5	-10.8	5	₩596,994,894	₩27,319,628	84	40470.27	2946.83	0.82
WCtyuy	-5.5	14	-3.5	20	₩517,073,770	-₩52,601,496	3	39289.90	5031.13	0.82

LN(ratio) = the sum of natural log relative error [$\ln(\text{error}_{HF}/\text{error}_{tsm})$] for MAD or RMSE of each forecasting method over the items in ME; HF-*tsm* = the total inventory costs of each hierarchical forecasting method deducted by the total inventory costs of *tsm*; Rank = the rank of the forecasting method in the 220 forecasting methods.

In Table 5-35, for Gun/RD, no forecasting method was identified as a robust forecasting method, although *SCtyum* was ranked within the top 10 for all the three measures [i.e. the sum of the LN(ratio) for MAD and RMSE, and the total inventory costs] with a qualified fill rate (i.e. greater than 70%). In Table 5-36, for ME, *SCtquy* was the most robust forecasting method and superior to *tsm* in terms of the three measures. In Table 5-37, for GE/AC, no forecasting method was ranked within the top 10 for all the three measures in common.

Table 5-37 Top 10 hierarchical forecasting methods for GE/AC in terms of either the natural log relative error for MAD or RMSE, or the total inventory costs

	MAD		RMSE		Simulation					
	LN(ratio)	Rank	LN(ratio)	Rank	Total inventory costs	HF-*tsm*	Rank	Mean Stock	Mean Backorder	Fill rate
TD1tsm	-3.0	8	-3.8	7	₩107,256,575	₩51,335,178	175	5867.59	344.21	0.83
TD2tm	3.6	48	2.3	44	₩46,816,337	-₩9,105,060	2	5941.18	612.06	0.69
TD2tsm	-3.3	6	-4.3	5	₩50,338,878	-₩5,582,519	11	5953.13	439.46	0.77
TD2tq	4.1	52	1.4	37	₩44,727,227	-₩11,194,170	1	5086.00	741.45	0.66
TD2uy	11.1	134	7.7	99	₩49,204,174	-₩6,717,223	8	8184.22	421.97	0.85
SCtmum	-3.4	5	-5.1	4	₩53,200,105	-₩2,721,292	25	7791.55	334.37	0.82
SCtmsm	-5.5	3	-5.9	3	₩55,212,827	-₩708,570	53	7092.00	288.35	0.84
SCtsmsm	-2.2	10	-2.4	11	₩58,415,518	₩2,494,121	83	7221.40	255.90	0.86
SCtqum	-6.6	1	-7.9	1	₩52,849,515	-₩3,071,882	22	7270.65	321.87	0.82
SCtqsm	-6.2	2	-6.6	2	₩54,315,965	-₩1,605,432	38	6606.75	301.30	0.83
SCtquq	-2.7	9	-3.1	10	₩50,791,881	-₩5,129,516	13	7581.80	308.32	0.82
SCtquy	-1.7	11	-3.7	8	₩124,453,833	₩68,532,436	184	6925.75	319.44	0.81
SCtsqum	-3.0	7	-4.2	6	₩57,142,233	₩1,220,836	77	7913.09	291.65	0.85
SCtsqsm	-3.5	4	-3.7	9	₩58,546,121	₩2,624,724	84	7251.27	278.70	0.85
SCuytm	13.2	147	12.5	154	₩47,446,316	-₩8,475,081	4	7319.22	379.08	0.76
SCuytsm	1.5	30	0.1	22	₩48,667,348	-₩7,254,049	6	7110.25	348.21	0.82
SCuytsq	4.7	57	4.0	58	₩46,896,551	-₩9,024,846	3	7221.84	370.36	0.78
SCtytm	9.9	116	10.6	138	₩50,272,102	-₩5,649,295	10	7162.12	538.28	0.68
WCuytm	7.5	80	8.1	101	₩47,919,716	-₩8,001,681	5	6719.71	542.79	0.68
WCuytsm	1.5	29	1.6	39	₩49,639,365	-₩6,282,032	9	6746.37	416.57	0.77
WCuytq	10.4	121	10.6	139	₩48,949,611	-₩6,971,786	7	6560.72	604.01	0.66

LN(ratio) = the sum of natural log relative error [ln($error_{HF}/error_{tsm}$)] for MAD or RMSE of each forecasting method over the items in GE/AC; HF-*tsm* = the total inventory costs of each hierarchical forecasting method deducted by the total inventory costs of *tsm*; Rank = the rank of the forecasting method in the 220 forecasting methods.

If the most robust forecasting method for all the 300 items (i.e. *SCtqum*) is used for forecasting spare parts for Gun/RD and GE/AC, and the most robust forecasting method for the spare parts for ME (i.e. *SCtquy*) is used for forecasting spare parts for ME, the total inventory costs for all the 300 items can be calculated as ₩667,184,679 (£340,922). These inventory costs are 3.8% smaller than the total inventory costs from using only *SCtqum* for forecasting all the 300 spare parts; that is, ₩693,601,747 (£352,619).

Table 5-38 The performance of hierarchical forecasting methods

	Absolute & relative measures	Derivative measure
Superior to the most robust DF	16.4% & 17.7% of HFs to *tsm* by LN(ratio) for MAD & RMSE	15.9% of HFs to *um* by total inventory costs
The most robust forecasting method	*SCtqum*	
Proration method	Top-21 by LN(ratio): 14 SCs, 6 WCs, 0 TD1, & 1 TD2 Top-21 by mean rank in MAD/RMSE: 15 SCs, 5 WCs, 0 TD1, & 1 TD2	Top-20 by total inventory costs: 12 SCs, 8 WCs, 0 TD1, & 0 TD2
The most frequently higher ranked DF for Top-21 or Top-20	Item level: m^1 & u^2 Item level combination: *um* Group level: m^1 & q^1; t^2& ts^2 Group level combination: *tq*	Item level: m^1& u^2 Item level combination: *um* Group level: m^1 & t^2
The most robust method for equipment group	ME: *SCtquy*	

DF = direct forecasting; HF = hierarchical forecasting; TD = top-down forecasting; SC = simple combination; WC = weighted combination; LN(ratio) = natural log relative error [$\ln(\text{error}_{HF}/\text{error}_{DF})$]; *1* = data aggregation method; *2* = data adjustment method.

Table 5-38 presents the major findings of this section; that is, the performance of hierarchical forecasting methods in terms of absolute, relative and derivative measures. *SCtqum* was demonstrated to be the most robust forecasting method in terms of the LN(HF/*tsm*) for MAD and RMSE, the total inventory costs, and the LN(HF/*tsm*) for total inventory costs in the years 2005 ~ 2007. Among the 4 proration methods, the proration methods for combinatorial forecasting, especially simple combination, dominated the top 21 and the top 20.

The frequencies of direct forecasting methods at group and item levels for the top 21 hierarchical forecasting methods in terms of the relative accuracy measure and the top 20 hierarchical forecasting methods in terms of the derivative measure were investigated. For the top 20 hierarchical forecasting methods, at group level, *m* as the data aggregation method and *t* as the data adjustment method were the most frequently higher ranked methods; at item level, *m* as the data aggregation method and *u* as the data adjustment method were the most frequently higher ranked methods. As the combination of the data aggregation and adjustment methods, at item level, *um* was found to be the most frequently higher ranked forecasting method. In addition, the performance of forecasting methods for spare parts for the three equipment groups (i.e. Gun/RD, ME, and GE/AC) was investigated.

5.6 Summary and Conclusion

This chapter compared a range of direct and hierarchical forecasting methods using absolute, relative and derivative measures.

5.6.1 Summary of findings

The major findings (i.e. superior forecasting methods) of this chapter are presented as shown in Table 5-39. In the period 2004 ~ 2007, the forecast with yearly aggregated data adjusted for linear trend (*ty*) presented the minimum mean rank in terms of MAD and RMSE. With the exception of 2004, the forecast with monthly aggregated data adjusted for linear trend and additive seasonality (*tsm*) [followed by the forecast with monthly aggregated unadjusted data (*um*)] presented the minimum mean rank in terms of MAD and RMSE. In the period 2005 ~ 2007 *tsm* was the most robust direct forecasting method among the 10 direct forecasting methods tested. The difference in the forecasting performance between the periods might be caused by the influence of the two peak points in 2002 and 2003 upon the forecasts in 2004. In terms of the total inventory costs *um* was found to be the most robust direct forecasting method (followed by *tsm*). *um* was also found to be the most frequently higher ranked item level direct forecasting method for the top 21 and 20 hierarchical forecasting methods. However, *tsm* minimised the mean rank in terms of the total inventory costs which is consistent with the results from the absolute measures.

Table 5-39 Summary of the performance of forecasting methods

		Absolute & relative measures	Derivative measure
DF	The most robust DF	*tsm*	Total inventory costs: *um* Mean rank for total inventory costs: *tsm*
DF	The most robust DF for equipment group	Gun/RD: *um*; ME & GE/AC: *tsm*	
HF	Superior to the most robust DF	16.4% & 17.7% of HFs to *tsm* by LN(ratio) for MAD & RMSE	15.9% of HFs to *um*
HF	The most robust method	*SCtqum*	
HF	Proration method	Top-21 by LN(ratio): 14 SCs, 6 WCs, 0 TD1, & 1 TD2 Top-21 by mean rank in MAD/RMSE: 15 SCs, 5 WCs, 0 TD1, & 1 TD2	Top-20 by total inventory costs: 12 SCs, 8 WCs, 0 TD1, & 0 TD2
HF	The most frequently higher ranked DF for Top-21 or Top-20	Item level: m^1 & u^2 Item level combination: *um* Group level: m^1 & q^1; t^2 & ts^2 Group level combination: *tq*	Item level: m^1 & u^2 Item level combination: *um* Group level: m^1 & t^2
HF	The most robust method for equipment group	ME: *SCtquy*	

DF = direct forecasting; HF = hierarchical forecasting; TD = top-down forecasting; SC = simple combination; WC = weighted combination; LN(ratio) = natural log relative error [ln($error_{HF}$/$error_{DF}$)]; *1* = data aggregation method; *2* = data adjustment method.

The most robust forecasting method among the forecasting methods tested for the 300 items was *SCtqum* in terms of the LN(HF/*tsm*) for MAD and RMSE, the total inventory costs, and the LN(HF/*tsm*) for total inventory costs in the years 2005 ~ 2007. As such, the internal validity of the robustness of *SCtqum* is claimed to be established. In the years 2005 ~ 2007, among the 220 hierarchical forecasting methods, 36 (16.4%), 39 (17.7%), and 35 (15.9%) of the hierarchical forecasting methods were superior to *tsm* in terms of the LN(ratio) for MAD and RMSE and the total inventory costs respectively. Combinatorial forecasting methods, especially simple combination methods dominated the top 21 and 20 hierarchical forecasting methods. The domination of combinatorial forecasting in the top 21 and 20 hierarchical forecasting methods corroborated the literature (Kahn, 1998, Dekker et al., 2004, Hyndman et al., 2007) in that combinatorial forecasting could present lower forecasting errors and lower inventory costs than top-down and direct forecasting. The domination of simple combination in the top 21 and 20 hierarchical forecasting methods corroborated DeLurgio (1998).

5.6.2 Forecasting scheme for the South Korean Navy

As stated earlier, the way that the South Korean Navy forecasts the spare parts demand is inappropriate because it does not capture the characteristics of the demand. A forecasting scheme for the South Korean Navy was derived from the results of this chapter as follows:

a) In lieu of the current direct forecasting, hierarchical forecasting is suggested.
b) As a proration method for hierarchical forecasting, combinatorial forecasting, especially simple combination, should be considered.
c) A careful selection of a forecasting method from various forecasting methods using simple combination is required, because the performance of the forecasting methods using simple combination for forecasting the 300 spare parts demand was highly variable.
d) As a forecasting method using simple combination, *SCtqum* is recommended for

forecasting spare parts demand for Gun/RD and GE/AC, because *SCtqum* generally provided the most robust forecasting performance for forecasting the 300 spare parts demand.

e) For forecasting spare parts demand for ME which is characterised as having lower correlation, steeper downward trend, lower intermittency, and larger demand volume, *SCtquy* is recommended, because *SCtquy* provided the most robust forecasting performance for forecasting the spare parts demand for ME.

f) Verification of a forecasting performance using simulation before implementing the forecast should be conducted. The simulation can reduce the risk of a wrong decision and guarantee the best practical decision in terms of monetary value and service level.

5.6.3 Conclusion

In this chapter, the performance of the direct and hierarchical forecasting methods was measured by the three groups of accuracy measures. With the three-fold measurements, reliability and internal validity of the results are claimed to be established. This chapter identified the robust forecasting methods and proposed the forecasting scheme for the South Korean Navy. Therefore, this chapter can claim to answer research question b) "what forecasting method is appropriate for the spare parts demand in the South Korean Navy?" by providing the robust forecasting methods. This also answers a part of research question c) "under what conditions are top-down forecasting or combinatorial forecasting superior or inferior to direct forecasting?" by providing the forecasting scheme.

The above forecasting scheme, which suggests using *SCtqum* for Gun/RD and GE/AC and *SCtquy* for ME, should be used with caution because it is based on the investigation with the equipment groups consisting of a small number of spare parts for the selected three types of warships. It might be difficult to apply these forecasting methods to forecasting other demands for the spare parts of other pieces of equipment or other types of warships in the South Korean Navy. The relationship between the demand features of ME and the forecasting performance of ME was suggested in this forecasting scheme. However, this forecasting scheme cannot explain the effect of the demand features upon the relative forecasting performance explicitly. In the next chapter, the effect of the demand features upon the relative forecasting performance will be investigated. A classification model which guides a choice of a superior forecasting strategy will be proposed.

Chapter 6. Forecasting Performance And Demand Features

In Chapter 4, the demand features of the spare parts in the South Korean Navy were analysed. In Chapter 5, the most robust direct forecasting method (i.e. *tsm*) and the most robust hierarchical forecasting method (i.e. *SCtqum*) for predicting the spare parts demand were identified in terms of the absolute, relative, and derivative measures. The robust forecasting method for spare parts for Main Engines was also identified. However, the forecasting scheme suggested in Subsection 5.6.2 cannot explicitly explain the relationships between the demand features of spare parts for the equipment groups and the performance of the forecasting methods for the spare parts for the equipment groups. This might have a limitation because those forecasting results for spare parts for Main Engines were derived from examinations with the limited number of spare parts. Hence, it might be difficult to apply the forecasting scheme to forecast other spare parts demands which have not been tested in the South Korean Navy. In this chapter, a classification model which uses demand features to predict the relative performance of alternative forecasting methods is examined, so that this model can be more generalisable than the forecasting scheme. A classification model to predict a superior forecasting method between the most robust hierarchical forecasting method and the most robust direct forecasting method by the multivariate demand features is proposed.

This chapter starts by describing the competing performance of the most robust direct forecasting method and the most robust hierarchical forecasting method in Section 6.1. In Section 6.2, possible demand features to guide the selection of a forecasting method are outlined. In Section 6.3, the process of classifying demands for the selection of a superior forecasting method by the features of the demands is described. In Section 6.4, classification results are related. Finally, a summary and concluding remarks are presented in Section 6.5.

6.1 Competing Performance

In Sections 5.4 and 5.5, a group of good forecasting methods were identified. Of the direct forecasting methods, *tsm* was generally found to be the most robust forecasting method for the 300 selected items in terms of MAD and RMSE, although *um* was found to provide a better result in terms of the total inventory costs. Of the hierarchical forecasting methods, *SCtqum* was generally found to be the most robust forecasting method for the 300 items in terms of most of the measures. *SCtqum* was unveiled as the superior forecasting method compared to the best direct forecasting method, *tsm*, in terms of all measurements. However, if the forecasting performance for each item is individually assessed, *SCtqum* is not always superior to *tsm* for each item.

Figure 6-1 compares the relative forecasting performance between the most robust direct forecasting method (i.e. *tsm*) and the most robust hierarchical forecasting method (i.e. *SCtqum*) for the 300 items between 2005 and 2007. Approximately, the share of the superiority for the 300 items is 40 % versus 60 % for *tsm* and *SCtqum* respectively. This implies that there is a 40 % chance of improvement compared with employing only *SCtqum* for the 300 items. If there is any guideline to select a forecasting method between *tsm* and *SCtqum*, a better performance than just employing a single forecasting method could be achieved.

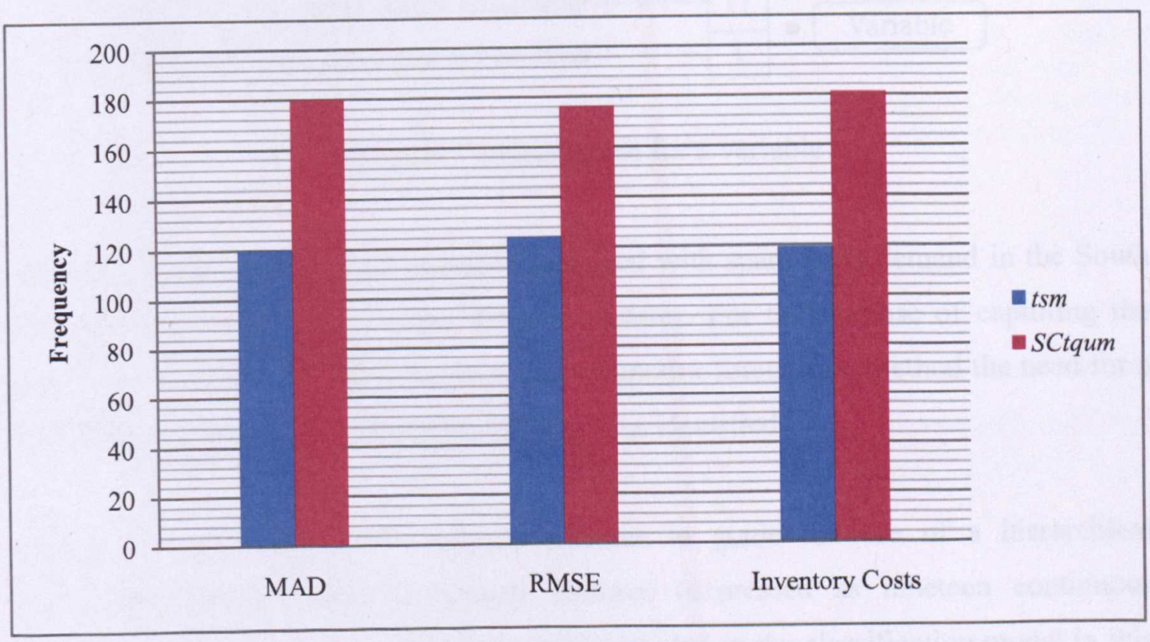


Figure 6-1 Relative forecasting performance (*tsm* vs. *SCtqum*)

6.2 Possible Demand Features to Guide the Selection of a Forecasting Method

Some guidelines for the use of top-down forecasting in previous research were reviewed in Subsection 2.5.2. The previous research is about guidelines for the use of top-down forecasting. However, these might also present an implication to guide the use of combinatorial forecasting because combinatorial forecasting can be considered to be a variant of top-down forecasting.

In this research, the most robust direct forecasting method (*tsm*) was generated by monthly aggregated time series at item level; the most robust hierarchical forecasting method (*SCtqum*) was generated by the simple combination of the forecast with quarterly aggregated time series at group level and the forecast with monthly aggregated time series at item level. Hence, the demand features for monthly aggregated time series at item level as well as the demand features for quarterly aggregated time series at group level were examined. In order to represent a variable, an abbreviation scheme was used as shown in Figure 6-2. For example, I.Slope denotes the Slope of monthly aggregated time series at item level; G.Cv(size) denotes the coefficient of variation in demand size in the quarterly aggregated time series at group level.

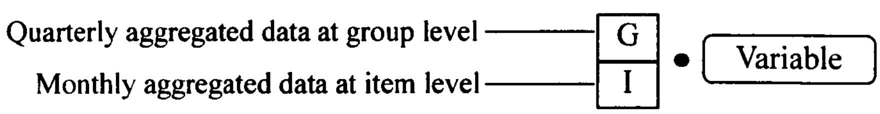


Figure 6-2 Abbreviation for a variable

As stated earlier, non-normal demand associated with spare parts demand in the South Korean Navy encompasses many demand features. For the purpose of capturing the nature of non-normal demand for selecting a superior forecasting method the need for a multidimensional calibration of data features was identified.

Among the various possible demand features to guide the use of a hierarchical forecasting method, eight continuous features (expressed as nineteen continuous variables) and two categorical features were included in the classification model in this research. Table 6-1 presents possible demand features and variables that might capture the demand features to guide the selection of a forecasting method with their impacts on the relative performance of forecasting strategies in research.

Table 6-1 Possible variables to guide the selection of a forecasting method

Demand feature	Variable	Impact on relative performance	Reference
Correlation	I.Corr(group) I.Corr(item)	↓ TD	Schwarzkopf et al. (1988)
		↓ DF (item)	Gross and Sohl (1990) & Widiarta et al.(2006)
		↑ DF (group) & BU	Fliedner (1999)
		Non-significance	Dangerfield and Morris (1992), Widiarta et al. (2008a) & Widiarta et al. (2009)
Variability	G.Cv(size) I.Cv(size)	↑ TD	Schwarzkopf et al. (1988)
	G.Pr(peak) I.Pr(peak)	↑ DF (group)	Viswanathan et al. (2008)
Forecasting horizon	PROLT	↑ TD	Shlifer and Wolff (1979)
UV	G.Mean I.Mean	↑ TD	Fliedner and Mabert (1992)
DV	Price×G.Mean Price×I.Mean		
Trend	G.Slope I.Slope	-	Businger and Read (1999)
Deviation from a normal distribution	G.Skewness I.Skewness	-	Businger and Read (1999)
	G.Kurtosis I.Kurtosis	-	-
Intermittency	G.Pr(zero) I.Pr(zero)	-	Johnston (1980), Johnston and Boylan (1996), Businger and Read (1999), Syntetos (2001), & Boylan et al. (2008)
Categorical variable	Year Equipment	-	-

UV: historical unit volume; DV: historical dollar volume; Corr(item or group) = correlations of the item level time series with other item level time series in the same group or the group level time series; Cv(size) = coefficient of variation in demand size; Pr(peak) = proportion of peak demands; PROLT = procurement lead time; Price = unit purchasing price; Pr(zero) = Proportion of zero demand periods; ↑ (or ↓) = increasing (or decreasing) the value of the demand feature increases the relative performance of the forecasting strategy; TD = top-down forecasting; BU = bottom-up forecasting; DF(group or item) = direct forecasting at group or item level.

Correlation

As discussed in Subsection 2.5.2, there was inconsistency about the impact of correlations. As shown in Subsection 4.3.3, in order to test the influence of correlations upon the relative performance of the alternative forecasting methods (*SCtqum* and *tsm*), the two kinds of correlations such as Corr(item) and Corr(group) were employed in the classification model of this research.

Variability

In order to measure the variability of non-normal demand, 'coefficient of variation in demand size' (Williams, 1984, Businger and Read, 1999, Syntetos, 2001) and 'number of peaks' (Businger and Read, 1999) were employed. As discussed in Subsection 2.5.3, there have been a few investigations about the influence of variability upon the performance of hierarchical forecasting. Schwarzkopf et al. (1988) examined the performance of direct and top-down forecasting methods at item level with missing or unreliable data. Viswanathan et al (2008) examined the influence of variability of demand size upon the performance of direct and bottom-up forecasting methods at group level. Cv(size) and Pr(peak) at both group and item levels identified in Subsection 4.3.4 were employed in the classification model of this research.

Forecasting horizon

As mentioned in Subsection 2.5.2, Shlifer and Wolff (1979) claimed that a top-down forecasting method was preferred to a direct forecasting method as the forecast goes further into the future. In order to test the influence of forecasting horizon upon the relative performance between *SCtqum* and *tsm*, PROLT was utilised in the classification model of this research.

UV and DV

Historical unit volume (UV) and historical dollar volume (DV) are grouping criteria which have been claimed to increase the accuracy of top-down forecasting significantly (Fliedner and Mabert, 1992). In this research, quarterly mean demand at group level (to test the influence of the quarterly mean demand upon *tq* at group level of *SCtqum*), and monthly mean demand at item level (to test the influence of the monthly mean demand upon *um* at item level of *SCtqum* and *tsm*) were employed as UV. In Section 4.3, DV was calculated as either "DV= sum of historical consumption for an item between Jan 2002 and Nov 2007 × item unit price" in order to understand the general demand feature of items or "DV = historical demand for an item per year × item unit price" in order to form the pair group. In this chapter, in order to test the discriminating influence of DV between quarterly demand at group level and monthly demand at item level upon the relative performance of the alternative forecasting methods in the classification model, "item unit purchasing prices multiplied by quarterly mean demand at group level" and "item unit purchasing prices multiplied by monthly mean demand at item level" were included in the model to test the influence of DV.

Trend and deviation from a normal distribution

Businger and Read (1999) used trend, seasonality, and skewness to classify demands for forecasting. As shown in Subsection 4.3.5, most of the seasonal effects of the spare parts data obtained from the Navy were found to be non-significant. Hence, seasonality was not considered to be a demand feature to be used in the classification model. Slope (to test the influence of the linear trend upon the relative performance of the alternative forecasting methods), skewness and kurtosis (to test the influence of the deviation from a normal distribution upon the relative performance of the alternative forecasting methods) at both group and item levels were utilised in the classification model of this research.

Intermittency

'Number of periods with zero demand' (Businger and Read, 1999, Boylan et al., 2008) and 'average inter-demand interval (ADI)' (Johnston, 1980, Johnston and Boylan, 1996, Syntetos, 2001, Boylan et al., 2008) were employed in order to measure the intermittency of a demand as stated. However, no research has been carried out about the influence of intermittency upon the performance of hierarchical forecasting. Pr(zero)s at both group level and item level were employed in the classification model of this research.

Categorical features

In categorical data observations are sorted into discrete mutually exclusive categories; in ordinal data objects are put into a rank order; and in interval data the intervals between adjacent values are equal (Howitt and Cramer, 2008, Field, 2009). As shown in Sections 5.4 and 5.5, the relative performance of forecasting methods were different in the different equipment groups (i.e. Gun/RD, ME or GE/AC) and also different year by year. The three forecasting years (i.e. 2005, 2006, and 2007) might not be either interval data or ordinal data. The intervals of the performance of forecasting methods between adjacent years were not equal; that is, a change in the performance of a forecasting method from 2005 to 2006 was not the same as the change in the performance of a forecasting method from 2006 to 2007. The performance of a forecasting method also was not put into a rank order between 2005 and 2007. The classification model included two categorical features, which are the three forecasting years and the three equipment groups; namely Year and Equipment respectively.

6.3 Process of Classification

This section describes the process of the classification modelling. This includes a detailed description of the forecasts and their performance for the modelling, classification methods, the building process of the model, the diagnostics of the model, the cross-validation process and a description of the final model.

6.3.1 Forecasting performance

For the purpose of investigating the classification model which predicts the superiority of the alternative forecasting methods (*tsm* vs. *SCtqum*) by the demand features, the three years (i.e. 2005, 2006 and 2007) of the demand forecasts and the corresponding demand features of the data which were used to produce these forecasts were examined. As the classification model is a prediction model, the demand features in the previous periods were utilised. For example, in order to test the forecasting accuracy of 2005 (or 2006), the forecast ranging from January 2005 (or 2006) to procurement lead time plus the review periods (12 months) was generated and measured. Then, the influence of demand features of the data ranging from January 2002 to December 2004 (or 2005) upon the forecasting performance was examined.

For the classification model, the same spare parts data, as used for forecasting in Chapter 5, were employed. In Section 5.4, it was shown that the forecasts in 2004 were highly influenced by the peak points in 2002 and 2003. In order to mitigate the erratic effect from the peak points of the years 2002 and 2003, the forecasts from the year 2005 to the year 2007 were investigated. Thus, 900 forecasts (i.e. the 300 items × the 3 years) were produced. Forecasting accuracy was measured until November 2007.

In Subsection 2.6.1, the two alternative effects of absolute measures of accuracy were reviewed. An absolute measure of accuracy using squared error such as RMSE places heavier weight on large errors than other methods; an absolute measure using absolute deviation such as MAD is less sensitive to outliers. In practice, the results measured using MAD and RMSE were found to be similar in Sections 5.4 and 5.5. Therefore, in this chapter, the performance of the two alternative forecasting methods was compared in terms of an error measure using only absolute deviation. In Subsection 2.6.1, in order to measure forecasting errors across a large amount of data, the error measures divided by means were introduced. As such, MAD/A and RMSE/A were used in Chapter 5.

In this chapter, the performance of each forecast was measured by the absolute deviation divided by the item's monthly mean consumption in order to avoid biasing effects of an item with large consumption as shown in equation (6-1). For example, the absolute deviation between the observed demand and the estimated demand of an item in 2005 (t = 1) was divided by the monthly mean demand for the item between 2002 (t-3, k = 3) and 2004 (t-1). The sum of the absolute deviations divided by monthly mean over the 300 items (N = 300) for the 3 years (n = 3) (i.e. 900 observations) is defined as in equation (6-2).

$$\textit{Absolute deviation divided by monthly mean} = \frac{|y_{i,t} - \hat{y}_{i,t}|}{\bar{y}_{i,t}} \tag{6-1}$$

$$\textit{Sum of absolute deviations divided by monthly mean} = \sum_{t=1}^{n}\sum_{i=1}^{N}\frac{|y_{i,t} - \hat{y}_{i,t}|}{\bar{y}_{i,t}} \tag{6-2}$$

where:

$y_{i,t}$ = the observed demand for item i at time t

$\hat{y}_{i,t}$ = the estimated demand for item i at time t

$\bar{y}_{i,t}$ = the monthly mean demand for item i between t-k and t-1

$$k = \begin{cases} 3, t = 1 \text{ (i.e. 2005)} \\ 4, t = 2 \text{ (i.e. 2006)} \\ 5, t = 3 \text{ (i.e. 2007)} \end{cases}$$

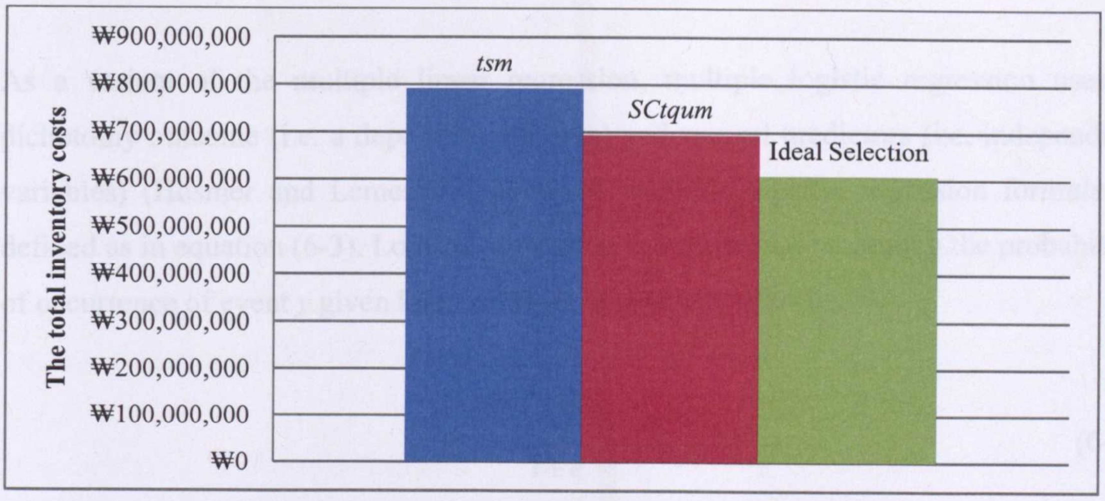
The sum of absolute deviations divided by the monthly mean over the 300 items for the 3 years for *SCtqum* was compared with those of the alternative forecasting methods. This included the top 21 hierarchical forecasting methods in terms of MAD and RMSE as in Subsection 5.5.2, and the alternative direct forecasting methods (i.e. *um* and *tsm*). Table 6-2 presents the sum of the absolute deviation divided by the monthly mean over the 900 observations for the alternative forecasting methods. Consistently, as with the results in Chapter 5, *tsm* presented smaller errors than *um*; *SCtqum* presented smaller errors than any other forecasting methods. This led to the choice of *SCtqum* and *tsm* as two alternative forecasting methods representing the most robust hierarchical and direct forecasting methods respectively in the classification models of this research.

Table 6-2 Forecasting performance comparisons in terms of the sum of absolute deviations divided by the monthly mean

Forecasting method	Sum of errors	Forecasting method	Sum of errors
um	8138.35	*tsm*	7877.95
TD2tm	12129.88	*SCtqsq*	7609.62
TD2tsm	7452.18	*SCtquy*	9191.39
TD2tq	8985.57	*SCtsqum*	7242.81
SCtmum	8818.46	*SCtsqsm*	7375.56
SCtmsm	8957.36	*SCtyum*	7430.90
SCtmuq	9036.49	*WCtmum*	8879.40
SCtsmum	7167.39	*WCtsmum*	7355.10
SCtsmsm	7291.10	*WCtqum*	7260.91
SCtsmuq	7474.08	*WCtqsm*	7523.94
SCtqum	7154.79	*WCtsqum*	7367.40
SCtqsm	7258.84	*WCtyum*	7612.62
SCtquq	7388.41		

Sum of errors = the sum of absolute deviations divided by the monthly mean over the 300 items for the 3 years.

Figure 6-3 compares the performance of *tsm*, *SCtqum* and an ideal selection. The total inventory costs and the sum of absolute deviations were calculated over the 300 items for the 3 years. The ideal selection represents a combination of superior forecasts between *tsm* and *SCtqum* for each observation in terms of equation (6-1). A series of simulations using the same simulation processes in Subsection 5.3.2 were conducted to calibrate the total inventory costs for *tsm*, *SCtqum* and the ideal selection. Considering that *SCtqum* is the most robust forecasting method, the ideal selection achieves considerable improvements over *SCtqum* in terms of the total inventory costs as well as the sum of errors. Building a classification model to approach the ideal selection might be a crucial way to achieve the improvement of forecasting accuracy so as to maximise the operational availability of weapon systems.



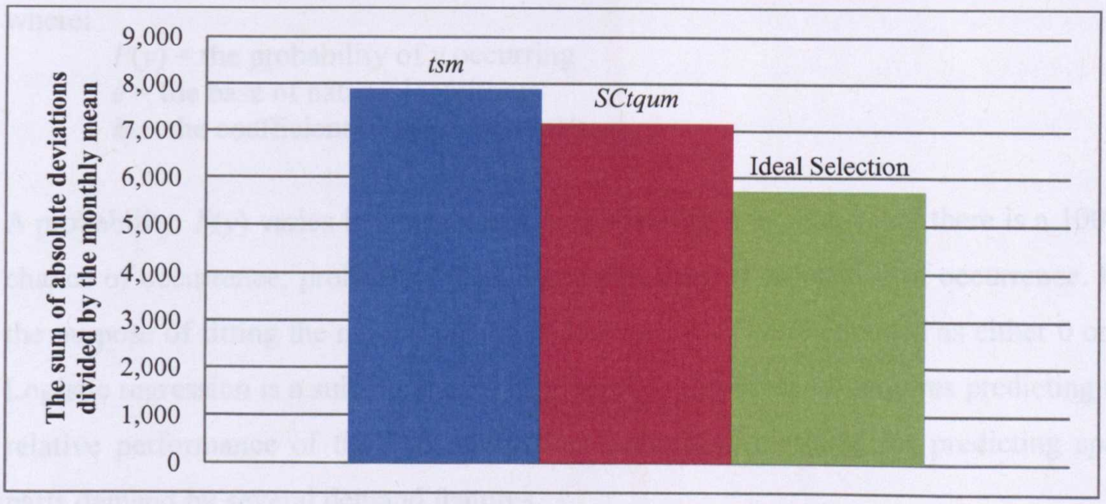


Figure 6-3 Forecasting performance comparisons

6.3.2 Various classification methods

Various classification methods such as multiple linear regression, logistic regression decision tree, and artificial neural networks to build a classification model were considered. As stated in Subsection 2.5.1, Businger and Read (1999) used multiple linear regression to examine the relationships between the accuracy of the four models such as three Box-Jenkins models [ARIMA(1, 1, 1), ARIMA(2, 2, 2) and ARIMA(3, 2, 3)] and exponential smoothing and the spare parts demand time series characterised by several statistics. However, they failed to find any distinct patterns in the relationships between the forecasting accuracy and the statistics. Linear regression cannot be used for dichotomous outcomes (i.e. *tsm* or *SCtqum*) because one of the assumptions of linear regression is that the relationship between outcome variables and predictors is linear (Field, 2009).

As a variant of the multiple linear regression, multiple logistic regression uses a dichotomy outcome (i.e. a dependent variable) and several predictors (i.e. independent variables) (Hosmer and Lemeshow, 2000). A multiple logistic regression formula is defined as in equation (6-3). Logistic regression is a model for predicting the probability of occurrence of event y given known values of x_j (Field, 2009).

$$P(y) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)}} \tag{6-3}$$

where:

$P(y)$ = the probability of y occurring
e = the base of natural logarithms
b_j = the coefficient of the j^{th} parameter, x_j

A probability, $P(y)$ varies between 0 and 1: probability 1 indicates that there is a 100 % chance of occurrence; probability 0 indicates that there is no chance of occurrence. For the purpose of fitting the model, the dependent variables were encoded as either 0 or 1. Logistic regression is a suitable model here because this research requires predicting the relative performance of the two alternative forecasting methods for predicting spare parts demand by several demand features.

Perlich et al. (2003) argued that a decision tree requires a large data set to present a better performance than logistic regression. They examined 36 large empirical two-class data sets (e.g. "Adult": whether income exceeds $50,000 per year; and "Bacteria": whether bacteria were found or not) which had at least 700 examples (i.e. observations). They argued that logistic regression was superior for smaller data sets and decision tree was superior for larger data sets. For example, in the case with the data set, Adult, decision tree was argued to outperform logistic regression for sets including more than around 10,000 observations in terms of classification accuracy (Perlich et al., 2003). A decision tree might be inappropriate for this research because only 900 observations (i.e. 900 annual forecasts for 300 items for 3 years) are available. Therefore, a decision tree was not considered for this research.

There are limitations of artificial neural networks (ANNs). Weights generated in ANNs are difficult to interpret. This difficulty of interpretation is called the "black box" nature of ANNs (Tu, 1996). Another limitation of ANNs is proneness to over-fitting (Ottenbacher et al., 2001). ANNs with several hidden layers may be able to achieve 100% accuracy by memorising all the cases; however, this over-fitted model cannot produce an accurate prediction with new data sets. Various methods to prevent over-fitting are available; however, some require losing information (Ottenbacher et al., 2001). Logistic regression is less likely to be over-fitted because the range of functions for logistic regression is more limited than ANNs (Ottenbacher et al., 2001).

Discrimination refers to a measure of how well the two classes in the data set are separated (Dreiseitl and Ohno-Machado, 2002). In practice there were no statistically significant difference in discrimination between artificial neural networks (ANNs) and logistic regression. Ottenbacher et al. (2001) postulated no statistically significant difference in discrimination between ANNs and logistic regression. Dreiseitl and Ohno-Machado (2002) analysed 72 academic papers and argued that in most of the statistical testing in discrimination, both ANNs and logistic regression performed at around the same level. Owing to the above limitations of ANNs and their similar performance to logistic regression, ANNs were not considered for this research.

6.3.3 Classification by logistic regression

In order to build the classification model which predicts the relative performance of the two alternative forecasting methods, multiple logistic regression was employed. "SPSS 15.0 for windows" was used to build a logistic regression model. When the performance of an *SCtqum* is superior to a *tsm* (i.e. the mentioned absolute deviations divided by the monthly mean of *SCtqum* was less than those of *tsm*) for an item, the dependent variable, $P(y)$ in equation (6-3), was encoded as 1; when the performance of a *tsm* is superior to an *SCtqum* (i.e. the mentioned absolute deviation divided by the monthly mean of *SCtqum* was greater than those of *tsm*) for an item, the dependent variable was encoded as 0. As such, the logistic regression model was fitted to the data obtained from the South Korean Navy.

In order to test the fitted logistic regression model, an outcome variable from the model was classified into either 0 or 1 with a cut-off value of c; that is, $P(y)$ larger than c was classified as *SCtqum*, whereas $P(y)$ smaller than c was classified as *tsm*. In this research, the cut-off value, 0.5 was employed because this is the most commonly used value for *c* (Hosmer and Lemeshow, 2000).

The above mentioned demand features of the data which were used for generating the forecasts were utilised as predictors. In logistic regression, predictors can be either continuous variables or categorical variables (Pallant, 2005). Nineteen continuous predictors and two categorical predictors which were shown in Table 6-1 were employed for this research. Dummy coding refers to a way of representing groups using only zeros and ones (Field, 2009). Two categorical predictors, Year and Equipment, were encoded as four dummy variables as in Table 6-3.

Table 6-3 Categorical variables coding

		Frequency	Parameter coding	
Equipment	Gun/RD	132	0	0
	ME	564	1	0
	GE/AC	204	0	1
Year	2005	300	0	0
	2006	300	1	0
	2007	300	0	1

In dummy coding, one group is considered to be a reference group (or a control group) against which all other groups should be compared (Miles and Shevlin, 2001, Field, 2009). Recalling the forecasting performance in each group, whilst *um* was the most robust direct forecasting method for Gun/RD, *tsm* was the most robust direct forecasting method for the other two groups. Recalling the demand features in each group, compared with other groups, Gun/RD was characterised as a more intermittent time series. Intermittency is an important demand feature which guided the selection of a forecasting method for non-normal demand (Johnston and Boylan, 1996, Boylan et al., 2008). Therefore, among the equipment groups, Gun/RD was determined to be the reference group in order to compare it with ME and GE/AC.

The forecasts in the year 2005 are more likely to be affected by the two peak points than the forecasts in the years 2006 and 2007. This is because the forecasts in the year 2005 are closer to the two peak points than the forecasts in the years 2006 and 2007. The year 2007, which might be least influenced by the two peak points, might not be appropriate to be used as the reference group because the year 2007 has an eleven-month period.

Therefore, the year 2005, which might be most influenced by the two peak points, was determined to be the reference group in order to compare the forecasts in 2005 with the forecasts in the years 2006 and 2007. As such, the reference groups were assigned to the code, '00'.

Selecting a predictor entry method together with identifying the predictors has a crucial impact on the performance of logistic regression (Field, 2009). Hierarchical predictor entry, forced entry and stepwise methods can be illustrated as predictor entry methods (Hosmer and Lemeshow, 2000, Miles and Shevlin, 2001, Field, 2009). In hierarchical predictor entry method, predictors are selected based on previous research and model builder's decision. These are entered into the model in order of their importance to the outcome (Field, 2009). Forced entry also stands upon existing theory about which predictors should be chosen, however, it forces all predictors into the model simultaneously without considering their importance (Field, 2009). Forced entry with all the 19 continuous predictors and the 2 categorical predictors was attempted for this research as follows.

In order to test the classification performance of the logistic regression model, the 10% cross-validation, which was stated in Subsection 3.6.2, was conducted. Among the 900 observations, 90 observations (i.e. 10% of the data set) were left out as a test set and 810 observations (90% of the data set) were used to build the model. This process was repeated 10 times, so as to test all the data sets. 10 logistic regression models based on forced entry method with the 10 training sets were built. However, the model found only 1 statistically significant (p-value < 0.05) predictor (i.e. Year) for 4 training sets; and no statistically significant predictor for 1 training set among the 10 training sets. The classification results from the models using forced entry method with either only 1 significant predictor (i.e. Year), which is not based on any existing theory, or no significant predictor were therefore potentially unreliable.

Stepwise method refers to a predictor entry method in which predictors are included in the regression model based on a statistical criterion (Hosmer and Lemeshow, 2000, Field, 2009). When the related theory is not well-developed, stepwise method is suggested (Hosmer and Lemeshow, 2000, Field, 2009). The stepwise process is continued until none of the remaining predictors have a significant score statistic (the cut-off point for p-value = 0.05) (Field, 2009). In lieu of forced entry method, stepwise method was considered. In this research, stepwise method is plausible. This is because there is little previous literature which verifies the significant influence of the demand features upon the performance of combinatorial forecasting; and a parsimonious model can be built using a stepwise method. A parsimonious model is defined as "one that explains the most variance in the dependent variable containing the fewest number of independent variables" (Miles and Shevlin, 2001, p. 38). This means that, all other things being equal, the simplest model is usually the best (DeLurgio, 1998).

Forward stepwise method starts with none of the predictors in the equation, then the most significant predictor is added into the equation (Miles and Shevlin, 2001, Field, 2009). On the contrary, in backward stepwise method, a computer starts calculations by placing all predictors in the equation, and assessing the contribution of each predictor to the outcome. Then, non-significant predictors are removed (Miles and Shevlin, 2001, Field, 2009). This research used both forward and backward stepwise methods, which could select different sets of predictors, to select predictors for the classification model.

Logistic regression estimates parameters (i.e. coefficients of the predictors) using maximum likelihood estimation. Maximum likelihood estimation maximizes the likelihood of obtaining the observed values of the dependent variable, given the independent variables (Miles and Shevlin, 2001). The log of the likelihood is defined as in equation (6-4). $P(y_{i,t})$ greater (or less) than 0.5 was classified as *SCtqum* (or *tsm*). The larger value of log-likelihood (*LL*), the better the model fits the data (Miles and Shevlin, 2001). The *LL* function is multiplied by -2, namely -2 log-likelihood (-2*LL*). This is for two reasons: a) to make *LL* a positive number because the *LL* is a negative number; b) to make *LL* distributed approximately as χ^2, in order to assess its significance with a χ^2 distribution (Miles and Shevlin, 2001). The smaller value of -2*LL*, the better the model fits the data, because the *LL* is multiplied by -2.

$$Log - likelihood = \sum_{t=1}^{n}\sum_{i=1}^{N}\{y_{i,t}\ln[P(y_{i,t})] + (1 - y_{i,t})\ln[1 - P(y_{i,t})]\} \tag{6-4}$$

where:

$y_{i,t}$ = the actual outcome of item i at time t
$y_{i,t}$ = 1 (or 0) indicates that *SCtqum* (or *tsm*) is superior
$P(y_{i,t})$ = the probability of $y_{i,t}$ for item i at time t (the cut-off value = 0.5)

When utilising the *LL* function different models can be compared as in equation (6-5) (Field, 2009). *LL*(Baseline) indicates the *LL* of a model which includes only the constant. The baseline model predicts all cases as the majority group (in this case, *SCtqum*). The χ^2 distribution has degree of freedom equal to the number of parameters in the new model minus the number of parameters in the baseline model (Field, 2009). Thus, the degree of freedom is merely the number of parameter in the new model minus one because the number of parameters in the baseline model is one (i.e. the constant).

$$\chi^2 = 2[LL(New) - LL(Baseline)] \tag{6-5}$$

As with linear regression, confidence limits are available for the j^{th} slope coefficient (b_j) and the intercept (b_0) of equation (6-3) (Hosmer and Lemeshow, 2000). 100(1-p)% point confidence limits for the estimated j^{th} slope coefficient ($\hat{b}_j$) can be calculated as in equation (6-6); those for the estimated intercept ($\hat{b}_0$) can be calculated as in equation (6-7).

$$\text{Confidence limits } (b_j) = \hat{b}_j \pm z_{\frac{1-p}{2}} \times \hat{SE}(\hat{b}_j) \tag{6-6}$$

$$\text{Confidence limits } (b_0) = \hat{b}_0 \pm z_{\frac{1-p}{2}} \times \hat{SE}(\hat{b}_0) \tag{6-7}$$

where:

p is the probability value for the confidence limits

$z_{\frac{1-p}{2}}$ is the upper $100(\frac{1-p}{2})$% point from the standard normal distribution

$\hat{SE}$ is the estimator of the standard error of the respective estimated coefficient

6.3.4 Diagnostics

Once a regression model is built, assessing model fit and the overall influence of an observation upon the model, namely the diagnostics of the model, is a critical issue (Field, 2009). A residual represents difference between an observed value and the value predicted by a model (Moore et al., 2009). Large residuals for a model imply a poorly fitted model and possible outliers. Standardized residuals refer to the residuals divided by an estimate of their standard deviation in order to standardise the scores to give them a mean of 0 and a standard deviation of 1 (Miles and Shevlin, 2001). Assuming standardized residuals have a normal distribution, 95% of the standardized residuals should lie between -1.96 and +1.96 and 99 % of the standardized residuals should lie between -2.58 and +2.58. Hence, in the process of building the classification model in this research, observations which lay outside of ±1.96 were removed.

Detecting improper influence of an observation upon the parameters of the model is also worthy of concern. Cook's distance is a measure of the overall influence of an observation upon the model (Cook and Weisberg, 1982). Cook and Weisberg (1982) claimed that a Cook's distance greater than 1 can introduce improper influence upon the parameters. Hence, in this research, observations scoring Cook's distance as greater than 1 were removed when constructing the model.

Multicollinearity is defined as "the size of correlations among the independent variables in a regression calculation" (Miles and Shevlin, 2001, p. 126). Multicollinearity can be introduced when predictors of a regression model are highly correlated (Pallant, 2005). Multicollinearity makes it difficult to estimate regression coefficient uniquely because there are large number of combinations of coefficients which work similarly (Field, 2009). Since each predictor contributes similar variance to the outcome, evaluating each predictor is also difficult (Miles and Shevlin, 2001, p. 126). Therefore, the regression coefficient would be unstable from data set to data set (Field, 2009).

Multicollinearity can be measured by a variance inflation factor (VIF) and tolerance (Miles and Shevlin, 2001, Field, 2009). VIF detects if a predictor has strong linear relationships with the other predictor(s). Tolerance indicates the reciprocal of VIF (i.e. 1/VIF). A tolerance smaller than 0.1 denotes a possible multicollinearity problem (Menard, 2002); VIF greater than 10 also can indicate a problem with multicollinearity (Myers, 1990). Diagnostics with the standardised residuals, Cook's distance, and multicollinearity were conducted for the logistic regression classification models in this research.

6.3.5 Cross-validation

As stated above, the 10% cross-validation was conducted for establishing internal validity of the classification model. Both forward and backward stepwise methods were employed for cross-validation. Table 6-4 presents predictors contained within each cross-validation set. The sets 1 ~ 10 indicate the sets for the 10% cross-validation which consisted of a 90% (810 observations: N = 270; n = 3) training set and a 10% (90 observations: N = 30; n = 3) test set from the total 900 observations (N = 300; n = 3). With the 10 pairs, all the data sets can be tested.

The equipment groups and the coefficient of variation in monthly demand size were identified as important demand features, because the two predictors [i.e. Equipment and I.Cv(size)] were contained within all sets as shown in Table 6-4. DV and Pr(zero) for both group and item level time series, G.Pr(peak) for group level time series, I.Skewness for item level time series, I.Corr(item), and PROLT were not contained within any set.

More item level demand features were contained than group level demand features. This might imply that demand features at group level have less effect on the relative performance of the alternative forecasting methods than demand features at item level. The time series data at group level were only used for generating a group level direct forecasting method (i.e. *tq*) for *SCtqum*; however, the time series data at item level were used for generating two item level direct forecasting methods (i.e. *um* and *tsm*) for *SCtqum* and *tsm* respectively. Hence, the item level demand features might influence more upon the relative performance of the two alternative forecasting methods than the group level demand features.

Table 6-4 Predictors contained within each cross-validation set

	Set	Sum	Category		Group level quarterly demand feature					Item level monthly demand feature					
			Year	Equipment	G.Slope	G.Cv(size)	G.Skewness	G.Kurtosis	G.Mean	I.Slope	I.Cv(size)	I.Pr(peak)	I.Kurtosis	I.Mean	I.Corr(group)
	1	4	1	1				1			1				
	2	4	1	1				1			1				
	3	4	1	1				1			1				
	4	2		1							1				
Forward stepwise method	5	6	1	1	1			1	1		1				
	6	2		1							1				
	7	2		1							1				
	8	4	1	1				1			1				
	9	2		1							1				
	10	2		1							1				
	Sum	32	5	10	1			5	1		10				
	1	4	1	1			1				1				
	2	6	1	1		1		1			1				1
	3	8	1	1						1	1	1	1	1	1
	4	8	1	1						1	1	1	1	1	1
Backward stepwise method	5	8	1	1						1	1	1	1	1	1
	6	6	1	1							1	1	1		1
	7	8	1	1						1	1	1	1	1	1
	8	9	1	1				1		1	1	1	1	1	1
	9	6	1	1		1		1		1	1				
	10	9	1	1	1		1				1	1	1	1	1
	Sum	72	10	10	1	2	2	3		6	10	7	7	6	8
Total		104	15	20	2	2	2	8	1	6	20	7	7	6	8

A suppressor variable is a variable having a very small correlation with the outcome variable. However due to a correlation with another predictor, it has a significant effect on the outcome variable (Conger, 1974, Lancaster, 1999). Backward stepwise method tends to include more suppressor variables than forward stepwise method (Field, 2009). Forward stepwise method is more inclined to exclude predictors involved in the suppressor effect than backward stepwise method. This is because the suppressor effect appears when a predictor has a significant effect but only when another predictor is held constant (Field, 2009).

As shown in the Table 6-4, the forward stepwise method was found to generate a more parsimonious model than the backward stepwise method. The forward stepwise method required fewer predictors than the backward stepwise method. This might be because forward stepwise method is more inclined to exclude predictors involved in the suppressor effect than backward stepwise method.

The performance of the foreward and backward stepwise methods in the data sets were measured by three measures. The sum of absolute deviations divided by the monthly mean over the items in the set for the 3 years [equation (6-2)], the total inventory costs from simulation, and Brier score (Brier, 1950) were employed. The Brier score can be expressed as in equation (6-8). $\hat{p}_{i,t}$ varies between 0 and 1. $\hat{p}_{i,t}$ greater (or less) than 0.5 was classified as *SCtqum* (or *tsm*). Steyerberg et al. (2001) intimated that the Brier score ranges from 0 (perfect) to 0.25 for sensible models: the bigger the score, the worse the quality of the prediction. The Brier score is suitable for quantifying overall accuracy of dichotomy predictions which is the case with the logistic regression model.

$$\text{Brier score} = \frac{1}{(n \times N)} \sum_{i=1}^{n} \sum_{t=1}^{N} (y_{i,t} - \hat{p}_{i,t})^2 \qquad (6\text{-}8)$$

where:

$y_{i,t}$ = the actual outcome of item i at time t

$y_{i,t}$ = 1 (or 0) indicates that *SCtqum* (or *tsm*) is superior

$\hat{p}_{i,t}$ = the forecast probability for item i at time t (the cut-off value = 0.5)

Table 6-5 The performance of backward and forward stepwise methods

			Training set			Test set		
		No. of predictors	Brier score	Errors	Costs(₩)	Brier score	Errors	Costs(₩)
Forward stepwise method	Set1	4	0.240	6,374	634,925,248	0.230	729	44,263,784
	Set2	4	0.236	6,233	620,367,219	0.259	886	64,292,651
	Set3	4	0.239	6,468	639,522,525	0.233	637	40,597,863
	Set4	2	0.243	6,173	638,408,750	0.230	837	37,478,283
	Set5	6	0.235	6,321	601,214,356	0.261	759	90,990,877
	Set6	2	0.240	6,452	600,620,770	0.260	581	75,434,066
	Set7	2	0.247	6,549	545,965,245	0.254	569	147,623,952
	Set8	4	0.238	6,265	542,286,696	0.250	861	142,473,586
	Set9	2	0.241	6,455	684,031,722	0.251	632	16,582,087
	Set10	2	0.241	6,337	660,116,966	0.251	736	20,613,099
	Sum		2.399	63,627	6,167,459,498	2.480	7,227	680,350,249
	Mean		0.240	6,363	616,745,950	0.248	723	68,035,025
Backward stepwise method	Set1	4	0.240	6,398	638,839,677	0.231	728	44,262,260
	Set2	6	0.235	6,168	620,720,318	0.264	930	65,484,473
	Set3	8	0.235	6,443	667,335,878	0.242	652	40,606,892
	Set4	8	0.237	6,241	653,881,485	0.227	807	38,140,787
	Set5	8	0.232	6,262	603,416,768	0.274	796	90,830,244
	Set6	6	0.236	6,512	605,500,089	0.252	566	77,820,417
	Set7	8	0.236	6,515	544,097,484	0.236	543	143,972,281
	Set8	9	0.233	6,189	555,778,834	0.259	865	139,693,982
	Set9	6	0.236	6,369	664,136,774	0.259	615	16,208,927
	Set10	9	0.234	6,303	665,544,084	0.265	731	21,055,107
	Sum		2.353	63,398	6,219,251,392	2.508	7,232	678,075,370
	Mean		0.235	6,340	621,925,139	0.251	723	67,807,537

Errors = the sum of the absolute deviations divided by the monthly mean over the set; Costs(₩) = the total inventory costs over the set; Sum (or Mean) = the sum (or mean) of the scores over the total training or test sets for forward or backward stepwise method.

Table 6-5 compares the performance of backward and forward stepwise methods in the 10% cross-validation sets. When comparing the performance in the training sets, backward stepwise method was superior in terms of the Brier Score and the errors; however, forward method was superior in terms of the costs. However, when comparing the performance in the test sets, the opposite results were observed. Forward stepwise method was superior in terms of the Brier score and the errors; however, the backward stepwise method was superior in terms of the costs. The superiority of forward stepwise method in terms of the Brier Score and the errors in test sets can be explained by the parsimonious model: that is, all other things being equal, the simplest model is usually the best.

A bias refers to the difference between estimated performance (i.e. the performance in the training data set) and test performance (i.e. the performance in the test data set).

Mean bias of the errors and mean bias of the costs were calculated as in equations (6-9) and (6-10) respectively.

$$Mean\ bias\ of\ errors = \left| \left\{ \frac{1}{(3\times 270)} \sum_{t=1}^{3} \sum_{i=1}^{270} \frac{|y_{i,t} - \hat{y}_{i,t}|}{\bar{y}_{i,t}} \right\}_{TR} - \left\{ \frac{1}{(3\times 30)} \sum_{t=1}^{3} \sum_{i=1}^{30} \frac{|y_{i,t} - \hat{y}_{i,t}|}{\bar{y}_{i,t}} \right\}_{TS} \right| \quad (6\text{-}9)$$

$$Mean\ bias\ of\ costs = \left| \left\{ \frac{1}{(3\times 270)} \sum_{t=1}^{3} \sum_{i=1}^{270} C_{i,t} \right\}_{TR} - \left\{ \frac{1}{(3\times 30)} \sum_{t=1}^{3} \sum_{i=1}^{30} C_{i,t} \right\}_{TS} \right| \quad (6\text{-}10)$$

where:

$y_{i,t}$ = the observed demand of item i at time t
$\hat{y}_{i,t}$ = the estimated demand of item i at time t
$\bar{y}_{i,t}$ = the monthly mean value of item i in the previous period
$C_{i,t}$ = the inventory costs of item i at time t
TR = training sets; TS = test sets

The superiority of forward stepwise method in terms of the Brier Score and the errors in the test sets can also be explained by the smaller bias for forward stepwise method. Table 6-6 compares mean biases between forward and backward stepwise methods for the cross-validation.

Table 6-6 Bias comparisons between forward and backward stepwise

Stepwise	Bias of Brier	Mean bias of errors	Mean bias of costs(₩)
Forward	0.008	0.175	5,470
Backward	0.016	0.208	14,932

It can be seen that forward stepwise method was less biased than backward stepwise method as shown. The smaller bias of forward stepwise method might cause the superiority of forward stepwise method in the test sets.

Table 6-7 presents the performance comparisons among the results in the test sets using forward stepwise method, using backward stepwise method, and without using classification model (i.e. only adopting *SCtqum*). It is interesting to note that, for both stepwise methods the classification results in the test sets presented lower total inventory costs than those of the case with only adopting *SCtqum*. However, in terms of the sum of the errors, both stepwise methods presented inferior performance to that of only adopting *SCtqum*.

Table 6-7 Performance comparisons among forward, backward and *SCtqum*

Stepwise	Sum of the errors	Total inventory costs(₩)
Forward	7,226.8	680,350,249
Backward	7,231.6	678,075,370
SCtqum	7,154.8	693,601,747

Sum of the errors = the sum of absolute deviations divided by the monthly mean over the 300 items for the 3 years; Total inventory costs (₩) = the total inventory costs (₩) over the 300 items for the 3 years.

Forward stepwise method was superior in terms of the sum of the errors; whereas backward stepwise method was superior in terms of the sum of the total inventory costs as shown in Table 6-7. However, when forward stepwise method was conducted with too small a number of predictors (i.e. 2), the performance of forward stepwise method was observed to have a tendency to be inferior to that of backward stepwise method in terms of the Brier score, the errors (i.e. the absolute deviation divided by the monthly mean), and the total inventory costs. 5 sets (i.e. the sets 4, 6, 7, 9, and 10) were identified as the cases with 2 predictors for forward stepwise method as shown in Table 6-5. As shown in Table 6-4, the forward stepwise methods for the sets 4, 6, 7, 9, and 10 selected the same 2 predictors such as I.Cv(size) and Equipment. When comparing the performance between forward and backward stepwise methods in the above 5 test sets, backward stepwise method was superior to forward stepwise method. The Brier score, mean errors per item, and mean total inventory costs per item for backward stepwise method in the above 5 test sets were 0.248, 7.25, and ₩660,439 (£337) respectively; those for forward method were 0.249, 7.45, and ₩661,626 (£338) respectively. As such, forward stepwise method with too small a number of predictors (i.e. 2) was identified to produce inferior performance than backward stepwise method for the 900 observations tested.

Therefore, a two step process for selecting a stepwise method was conducted. At the first step, for the purpose of achieving a parsimonious model, forward stepwise method is preferentially considered. At the second step, if forward stepwise method selects too small a number of predictors (i.e. 2), backward stepwise method is employed; otherwise, forward stepwise method is employed. This resulted in, sets 1, 2, 3, 5 and 8 employing forward stepwise method, whilst sets 4, 6, 7, 9 and 10 which included 2 predictors for forward stepwise method employing backward stepwise method.

Table 6-8 compares mean biases among forward stepwise, backward stepwise and the employed model. The employed model represents the employed logistic regression model combining the two stepwise methods using the above two step process (i.e. forward stepwise method for sets 1, 2, 3, 5 and 8; backward stepwise method sets 4, 6, 7, 9 and 10). The employed model presented the smallest bias in the errors. However, the bias of the Brier score and the mean bias of the costs for the employed model were in between forward stepwise method and backward stepwise method. This is logical as the employed model is a combination of backward stepwise method and forward stepwise method.

Table 6-8 Bias comparisons among forward, backward and employed

Stepwise	Bias of Brier	Mean bias of errors	Mean bias of costs(₩)
Forward	0.008	0.175	5,470
Backward	0.016	0.208	14,932
Employed	0.011	0.007	6,063

Mean bias of errors and mean bias of costs were calculated as in equations (6-9) and (6-10) respectively.

Table 6-9 presents the classification results using the employed model in the test sets compared with the results using only the most robust forecasting method (i.e. *SCtqum*). The mean Brier score of the classification model was 0.247 which is within the range for sensible models (Steyerberg et al., 2001). In terms of mean errors and mean costs, the performance of the model was observed to be marginally superior to that of *SCtqum*. In seven test sets out of all ten test sets, the model presented smaller errors than *SCtqum*. In six test sets out of all ten test sets, the model presented smaller costs than *SCtqum*.

Table 6-9 Classification results using the employed model

	Brier score	Errors		Costs(₩)	
	Model	*SCtqum*	Model	*SCtqum*	Model
Set1	0.230	733	**729**	50,848,742	**44,263,784**
Set2	0.259	911	**886**	65,439,615	**64,292,651**
Set3	0.233	668	**637**	41,355,311	**40,597,863**
Set4	0.227	834	**807**	**37,763,001**	38,140,787
Set5	0.261	764	**759**	91,462,305	**90,990,877**
Set6	0.252	598	**566**	84,190,734	**77,820,417**
Set7	0.236	569	**543**	147,623,952	**143,972,281**
Set8	0.250	**810**	861	**138,875,307**	142,473,586
Set9	0.259	**578**	615	**15,721,017**	16,208,927
Set10	0.265	**690**	731	**20,321,763**	21,055,107
Sum	2.471	7,155	**7,133**	693,601,747	**679,816,280**
Mean	0.247	715	**713**	69,360,175	**67,981,628**

Model = the classification model using the employed model; Errors = the sum of absolute deviations divided by the monthly mean over the set; Costs (₩) = the total inventory costs over the set; lower errors and lower costs are shown in bold.

Table 6-10 compares the prediction results classifying all 900 observations into *SCtqum* and the prediction results classifying the 900 observations according to the employed model in the cross-validation test sets. "Predicted" indicates these prediction results. "Observed" indicates the classification results from the ideal selection stated in Subsection 6.3.1. "Observed" classification results were compared with the "predicted" classification results. If the "observed" classification result for an observation is equivalent to the "predicted" classification result for the observation, this is counted as a correct prediction. While the total percentage of correctness using only *SCtqum* was 54.0, the total percentage of correctness using the employed model was 55.9%.

Table 6-10 Prediction results

	Observed	Predicted		Correct (%)
		tsm	*SCtqum*	
SCtqum	*tsm*	0	414	0
	SCtqum	0	486	100
	Overall percentage			54.0
Employed model	*tsm*	160	254	38.6
	SCtqum	143	343	70.6
	Overall percentage			55.9

Correct (%) = the percentage of correct predictions.

There was a tendency to classify more into the major class (i.e. *SCtqum*) in the predicted classification than in the observed classification. Whilst 486 (54.0%) observations were classified into *SCtqum* in the observed classification, 597 (66.3%) observations were classified into *SCtqum* in the predicted classification. Whilst 414 (46.0%) observations were classified into *tsm* in the observed classification, 303 (33.7%) observations were classified into *tsm* in the predicted classification. Hence, higher correctness (70.6%) for *SCtqum* was observed than correctness (38.6%) for *tsm* in the predicted classification.

The improvement of the forecasting performance of the employed model using the employed model in the test sets can be observed in Figure 6-4. The total inventory costs of the employed model in the test sets were calculated from the simulation using the predicted classification results of the model utilising the employed model. The sum of absolute deviations divided by the monthly mean as well as the total inventory costs decreased compared with those of *SCtqum*. The sum of the errors reduced from 7154.8 to 7132.9; the total inventory costs reduced from ₩693,601,747 (£354,420) to ₩679,816,280 (£347,377). Based on these results, it might be suggested that the internal validity is established for the classification model.

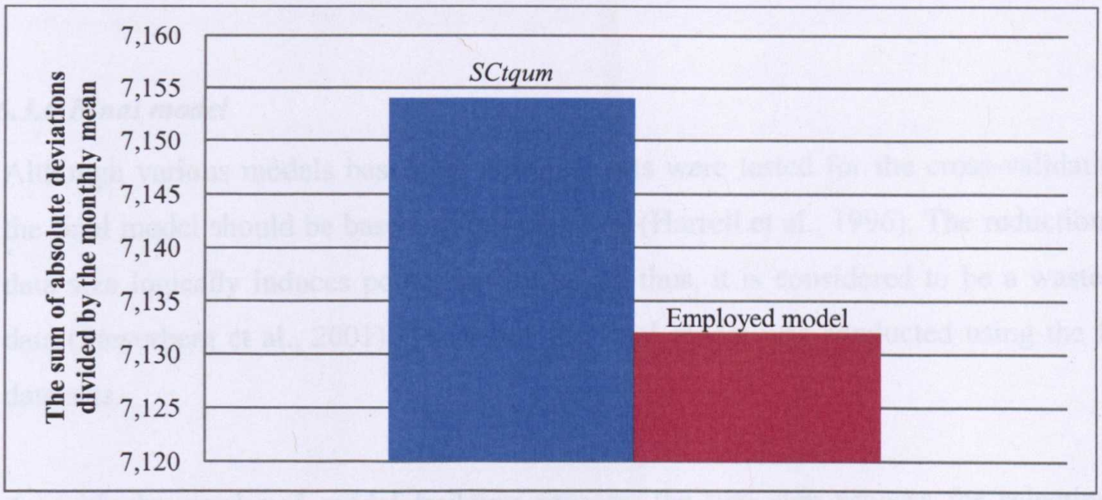


Figure 6-4 Forecasting performance improvement of the employed model

As stated in Subsection 3.6.3, internal validity can be seen as an approximation to external validity (Steyerberg et al., 2001). Therefore, the applicability of the classification model to other data sets in the South Korean Navy could be approximated. Since existing theories provide a framework to compare the results of this research, when more cases appear to support this classification model, replication can be claimed (McCutcheon and Meredith, 1993, Yin, 2003).

These results need to be interpreted with caution. The purpose of the classification is to select a superior forecasting method. Top-down forecasting was observed not to be a robust forecasting method in Section 5.5. Thus, top-down forecasting was not considered in the classification model. As this classification model compares a combinatorial forecasting method (i.e. *SCtqum*) and a direct forecasting method (*tsm*), it could not be asserted firmly that any result of this model is either consistent or inconsistent with previous research which has compared top-down forecasting and direct forecasting. As stated in Subsection 2.5.3, there has been no controlled research which examined the influence of demand features upon the performance of combinatorial forecasting. All of the previous researchers examined the influence of demand features upon the performance of top-down forecasting. This is a research gap as well as a contribution of this research. Nevertheless, the results of this research are worth comparing to previous research because combinatorial forecasting can be considered to be a variant of top-down forecasting. Likewise, existing theories could provide a framework for the findings of this research.

6.3.6 Final model

Although various models based on split data sets were tested for the cross-validation, the final model should be based on full data sets (Harrell et al., 1996). The reduction in data size logically induces poorer performance; thus, it is considered to be a waste of data (Steyerberg et al., 2001). Therefore, the final model was conducted using the full data sets.

As with the employed model building process, the two step process for selecting a stepwise method was conducted for the final model. At the first step, for the purpose of achieving parsimonious model, forward stepwise method was considered. At the second step, however, forward stepwise method was observed to select too small a number of predictors (i.e. two predictors). Therefore, backward stepwise method was employed for the final model.

The diagnostics of the final model was conducted. In the final model, the standardized residuals of all the observations lay inside of ± 1.96. Therefore, the model fitting was satisfactory in terms of standardized residuals. As mentioned above, a Cook's distance greater than 1 can introduce improper influence upon the parameters (Cook and Weisberg, 1982). In the final model, there was one observation greater than 1 (i.e. 1.32853). This observation is the item, Gasket1 in group 83, which is included in the equipment group, ME, in the years 2002 ~ 2004 for demand features (i.e. in the year 2005 for forecasting). It should be recognized that the final model could be biased by this one observation. If this observation is a substantial case, removing this case is not justified. A close inspection is required to eliminate such observations (Field, 2009). As stated in Section 4.5, there are three sources of distortion in the demand data with the South Korean Navy (i.e. the multi-echelon inventory systems, the budgeting process and the maintenance system). Unusually high demands for Gasket1 (i.e. more than 10 times higher than the monthly mean demand for Gasket1) were identified in December 2002 and December 2003. This might be caused by one of the three sources distorting the true demand. As such, this observation was removed, and then a new model was built without the observation. However, there was then another observation greater than 1 (i.e. 1.03692). This observation is the item, Sealing Ring in group 85, in the equipment group, ME, in the years 2002 ~ 2004 for demand features (i.e. in the year 2005 for forecasting). An unusually high demand for Sealing Ring (i.e. 7.5 times higher than the monthly mean demand for Sealing Ring) was identified in February 2004. This might also caused by one of the three sources of distortion. Therefore, a final logistic regression model without these two observations was built.

Table 6-11 presents Multicollinearity statistics for the predictors included in the final model. As shown in the table, all the tolerance statistics were greater than 0.1; and all the variance inflation factor (VIF) were smaller than 10. Therefore, there was no serious concern of multicollinearity for the final model.

Table 6-11 Multicollinearity statistics

Predictors	Tolerance	VIF
Year	0.891	1.123
Equipment	0.832	1.202
G.Kurtosis	0.560	1.787
I.Slope	0.372	2.689
I.Cv(size)	0.197	5.083
I.Corr(group)	0.777	1.287
I.Pr(peak)	0.440	2.272
I.Kurtosis	0.203	4.933
I.Mean	0.383	2.613

Tolerance < 0.1 (or 10 < VIF): a possible multicollinearity problem.

Overall, twelve steps of iterations using backward stepwise method were conducted for the final model. -2*LL* of the baseline model and the new model were 1241.899 and 1192.313 respectively. Thus, the χ^2 was 49.586 by equation (6-5). The number of parameters in the new model was 12; that is, 1 constant, 2 categorical predictors (i.e.

Year and Equipment) expressed as 4 dummy variables, and 7 continuous predictors (i.e. G.Kurtosis, I.Slope, I.CV(size), I.Corr(group), I.Pr(peak) I.Kurtosis and I.Mean). Thus, the degrees of freedom were 11. The new model was significant as the p-value was less than 0.001 by the χ^2 distribution.

6.4 Classification Results

After twelve steps of calculation by backward stepwise method using a likelihood ratio statistic, the logistic regression model produced the final model. This section includes the performance of the classification, interpretation of the models and the predictors, and also individual relations between the performance of the alternative forecasting methods and significant predictors.

6.4.1 Classification by the logistic regression model

The results of the final model are presented in Table 6-12. As stated in Subsection 3.6.2, resubstitution refers to an estimate which uses a data set to build the model as well as to test the model (White and Liu, 1997). "Observed" indicates the classification results from the ideal selection stated in Subsection 6.3.1. "Predicted" indicates the classification results produced using the final model. If the "observed" classification result for an observation is equivalent to the "predicted" classification result for the observation, this is counted as a correct prediction.

Table 6-12 Prediction results from resubstitution

Observed	Predicted		Correct (%)
	tsm	*SCtqum*	
tsm	188	226	45.4
SCtqum	139	347	71.4
Overall percentage			59.4

Correct (%) = the percentage of correct predictions using the classification model in the total observations.

As these results were from resubstitution, these presented an overly optimistic view of the true accuracy of the model, compared with the results from the employed model in the cross-validation test sets as shown in Table 6-10. The total percentage of correctness (59.4%) of resubstitution was higher than that (55.9%) with the test sets.

As with the classification results from the test sets, there was a tendency to classify more into the major class (i.e. *SCtqum*) in the predicted classification than in the observed classification. Whilst 486 (54.0%) observations were classified into *SCtqum* in the observed classification, 573 (63.7%) observations were classified into *SCtqum* in the predicted classification. Whilst 414 (46.0%) observations were classified into *tsm* in the observed classification, 327 (36.3%) observations were classified into *tsm* in the predicted classification. Hence, higher correctness (71.4%) for *SCtqum* was observed than correctness (45.4%) for *tsm* in the predicted classification. The two outlier items which had been screened out earlier for building the model were included to calculate the correctness of the final model. As it can be expected, the final regression model predicted wrong outcomes for theses two outlier items.

Figure 6-5 presents the comparisons of the forecasting performance among *SCtqum*, the employed model for the test sets and the final model for resubstitution in terms of the sum of absolute deviations divided by the monthly mean and the total inventory costs.

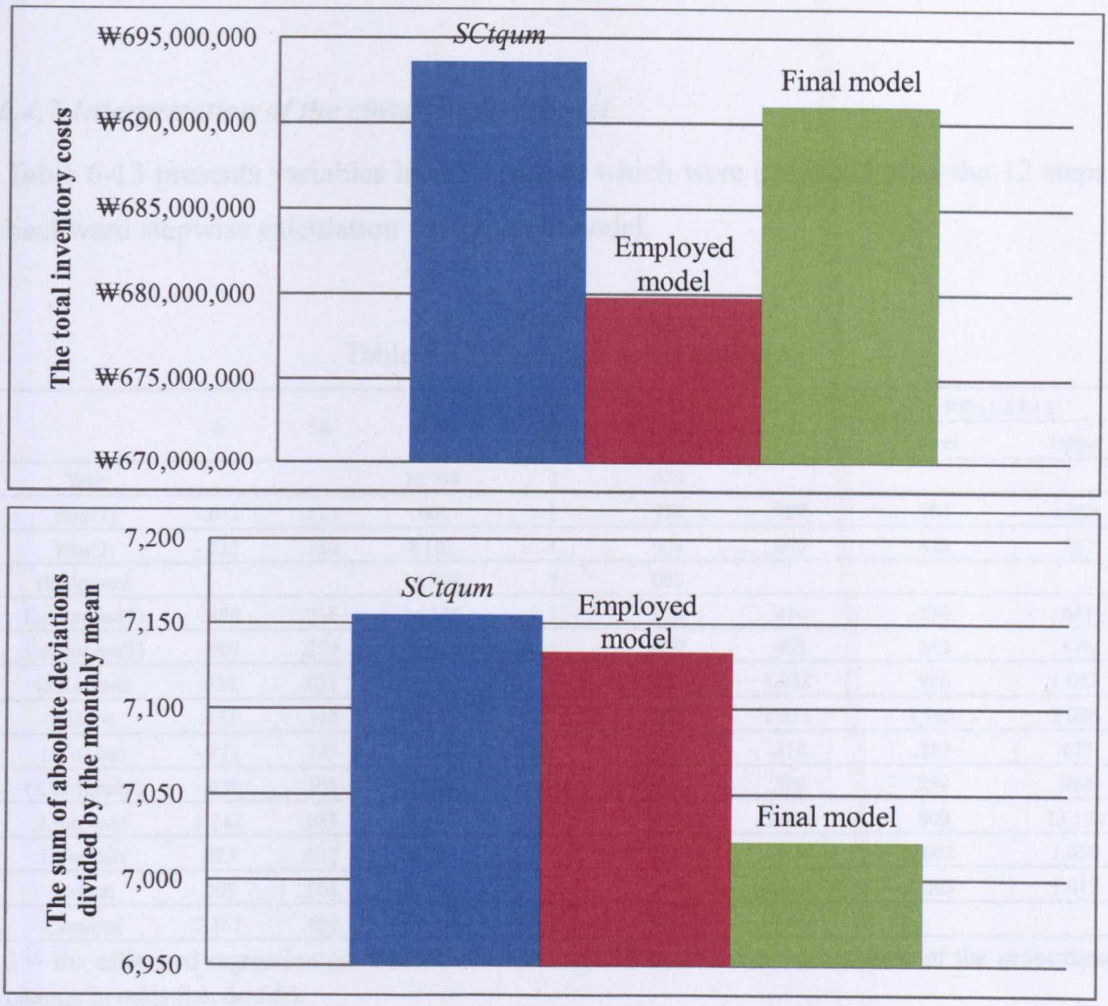


Figure 6-5 Forecasting performance improvement in the model

As resubstitution presents an optimistic view, the sum of forecasting errors were observed to reduce in order of *SCtqum*, the employed model and the final model. What was surprising is that the total inventory costs of the employed model were smaller than those of the final model; that is, the performance of the employed model in the split data was superior to the performance of the final model in the resubstitution. The smaller biases of forward stepwise method in the Brier score, the absolute deviation divided by the monthly mean and the total inventory costs than those of backward stepwise method were identified as shown in Table 6-6. It may be that the smaller bias of forward stepwise method in terms of the total inventory costs could create a superior performance to the performance of the final model which employs backward stepwise method. Remembering that both forward and backward stepwise methods were employed for the test sets, the employed model for the test sets can present the smaller total inventory costs than the final model.

6.4.2 Interpretation of the classification model

Table 6-13 presents variables in the equation which were produced after the 12 steps of backward stepwise calculation for the final model.

Table 6-13 Variables in the equation

	b	SE	Wald	df	Sig.	e^b	95% CI for e^b	
							Lower	Upper
Year			10.798	2	.005			
Year(1)	-.013	.173	.006	1	.938	.987	.704	1.384
Year(2)	-.512	.180	8.102	1	.004	.600	.422	.853
Equipment			16.974	2	.000			
Equipment(1)	-.870	.216	16.148	1	.000	.419	.274	.641
Equipment(2)	-.909	.259	12.314	1	.000	.403	.243	.670
G.Kurtosis	.038	.021	3.134	1	.077	1.038	.996	1.082
I.Slope	.452	.145	9.749	1	.002	1.571	1.183	2.086
I.Cv(size)	-.872	.245	12.695	1	.000	.418	.259	.675
I.Corr(group)	-.639	.307	4.327	1	.038	.528	.289	.964
I.Pr(peak)	1.242	.683	3.312	1	.069	3.463	.909	13.194
I.Kurtosis	.025	.012	4.391	1	.036	1.026	1.002	1.050
I.Mean	.009	.004	7.163	1	.007	1.010	1.003	1.017
Constant	2.367	.452	27.447	1	.000	10.664		

b = the estimated regression coefficient; SE = the standard error; e^b = an indicator of the proportionate change in odds (i.e. $\Delta odds$).

The Wald statistic tests whether the b-coefficient in equation (6-3) for a predictor is significantly different from zero (Field, 2009). When the b-coefficient is significantly different from zero, the b-coefficient is implied to contribute significantly to the outcome. The Wald statistic can be calculated as in equation (6-11) (Hosmer and Lemeshow, 2000).

$$Wald = \left(\frac{b}{SE_b} \right)^2 \qquad (6\text{-}11)$$

Odds (i.e. the probability of an event occurring divided by the probability of that event not occurring) are defined as in equation (6-12) (Miles and Shevlin, 2001, p. 155). The odds ratio, $\Delta odds$ is defined as in equation (6-13) (Field, 2009, p. 271). The relationship between $\Delta odds$ and the b-coefficient is represented as in equation (6-14) (Hosmer and Lemeshow, 2000).

$$Odds = \frac{P(y)}{1 - P(y)} \tag{6-12}$$

where: $P(y) = \dfrac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)}}$

$$\Delta Odds = \frac{odds\ after\ a\ unit\ change\ in\ the\ predictor}{odds\ before\ change} \tag{6-13}$$

$$\Delta Odds = e^{b} \tag{6-14}$$

A value, e^b (i.e. Δ*odds*) of greater than 1 denotes that, when the predictor increases, the odds of the outcome occurring increase; and a value, e^b of less than 1 denotes that, when the predictor increases, the odds of the outcome occurring decrease. In this research, *tsm* was encoded as 0 and *SCtqum* was encoded as 1. Thus, the e^b of greater than 1 indicates that: as the predictor increases, the odds of classifying *SCtqum* increase. The e^b of less than 1 indicates that: as the predictor increases, the odds of classifying *tsm* increase.

100(1-*p*)% point confidence limits for the estimated coefficient ($\hat{b}_j$) was presented as in equation (6-6). The corresponding 100(1-*p*)% point confidence limits for the Δ*odds* (i.e. e^b) can be calculated as in equation (6-15) (Hosmer and Lemeshow, 2000). 95% confidence intervals (CI) for the e^b are presented as shown in Table 6-13.

$$\text{Confidence limits } (e^{b}) = e^{\left[\hat{b}_j \pm z_{\frac{1-p}{2}} \times \hat{SE}(\hat{b}_j)\right]} \tag{6-15}$$

where:

p is the probability value for the confidence limits

$z_{\frac{1-p}{2}}$ is the upper $100(\frac{1-p}{2})$% point from the standard normal distribution

$\hat{SE}$ is the estimator of the standard error of the respective estimated coefficient

In the final model, five continuous predictors [i.e. I.Slope, I.Cv(size), I.Corr(group), I.Kurtosis and I.Mean] and one categorical predictor (i.e. Equipment) were significantly different from zero according to the Wald statistic. As I.Slope, I.Kurtosis or I.Mean increases, the odds of classifying *SCtqum* increase in terms of e^b. On the contrary, as I.Cv(size) or I.Corr(group) increases, the odds of classifying *tsm* increase in terms of e^b.

For the 10% cross-validation, both forward and backward stepwise methods used 90 % of the data set (i.e. 810 observations) to build a model and were repeated 10 times respectively. In addition to the significant predictors for the final model, other predictors were also observed to be significant for the classification models using both forward and backward stepwise methods for the 10% cross-validation. In forward stepwise method, among the predictors contained within each set (as shown in Table 6-4), G.Slope, G.Kurtosis and G.Mean were additionally observed to be significant as the *p*-values of the Wald statistic, as shown in equation (6-11), were less than 0.05. In backward stepwise method, among the predictors contained within each set (as shown in Table 6-4), G.Slope, G.Kurtosis and I.Pr(peak) were additionally observed to be significant for the models as the *p*-values were less than 0.05. The e^b scores of all of these additionally significant predictors were observed to be greater than one. This means that as G.Slope, G.Mean, G.Kurtosis or I.Pr(peak) increases, the odds of classifying *SCtqum* increase.

6.4.3 Predictors in the classification model

Among the 19 continuous predictors and the 2 categorical predictors, the 7 continuous predictors and the 2 categorical predictors were included in the final model. Among them, only the 5 continuous predictors and the 1 categorical predictor were significant as shown in Table 6-13. Additionally, 4 more predictors were found to be significant in terms of the Wald statistic in the employed classification models for the 10% cross-validation. Table 6-14 presents the relations between the 10 significant predictors and odds of classification. Related references to the findings of this research are presented in the last column. These are discussed as follows.

Table 6-14 Relations between predictors and odds of classification in the models

Demand feature	Significant predictor	Odds of classification	Reference
Correlation	I.Corr(group)	I.Corr(group) → *tsm*	= Schwarzkopf et al. (1988); ≠ Gross and Sohl (1990), Dangerfield and Morris (1992), Fliedner (1999), Widiarta et al.(2006), Widiarta et al. (2008a), & Widiarta et al. (2009)
Variability	I.Cv(size) I.Pr(peak)	I.Cv(size) (or I.Pr(peak)) → *tsm* (or *SCtqum*)	≠ (or =) Schwarzkopf et al. (1988) & Viswanathan et al. (2008)
Forecasting horizon	Non-sig.	-	≠ Shlifer and Wolff (1979)
UV & DV	G.Mean I.Mean.	G.Mean or I.Mean → *SCtqum*	= Fliedner and Mabert (1992)
Trend	G.Slope I.Slope	G.Slope or I.Slope → *SCtqum*	-
Deviation from a normal distribution	G.Kurtosis I.Kurtosis	G.Kurtosis or I.Kurtosis → *SCtqum*	-
Intermittency	Non-sig.	-	-
Categorical variable	Equipment	The odds of classification in Gun/RD are significantly different from the odds in ME and GE/AC	-

→: increasing the value of the predictor increases the odds of classifying the forecasting method;
= (or ≠): consistent (or inconsistent) with the finding of the research; Non-sig.: non-significance;
UV: historical unit volume; DV: historical dollar volume.

Correlation

Between the statistics [I.Corr(item) and I.Corr(group)] representing correlations, I.Corr(group) was observed to be a significant predictor for the final model and some of the models using backward stepwise method for the 10% cross-validation. By the e^b, the odds of classifying *tsm* increase with the increasing I.Corr(group). This result might be consistent with the analytic argument of Schwarzkopf et al. (1988). Schwarzkopf et al. (1988) contended that, when two item level time series are independent, forecasting errors for direct forecasting are more variable than forecasting errors for top-down forecasting; however, when there are strong positive correlations between the items, the sum of the variability of the top-down forecasting errors is greater than the sum of the variability of the direct forecasting errors. Hence, this result might be inconsistent with the research which stands for the opposite argument (Gross and Sohl, 1990, Fliedner, 1999, Widiarta et al., 2006) and the research which claimed the non-significance of the correlations (Dangerfield and Morris, 1992, Widiarta et al., 2008a, Widiarta et al., 2009).

Variability

Among the statistics [G.Cv(size), I.Cv(size), G.Pr(peak), and I.Pr(peak)] representing the variability of demand size, I.Cv(size) was observed to be a significant predictor for the final model, and I.Pr(peak) was observed to be a significant predictor for some of the models using backward stepwise method for the 10% cross-validation. By the e^b, the odds of classifying *tsm* increase with the increasing I.Cv(size); however, the odds of classifying *SCtqum* increase with the increasing I.Pr(peak).

The case with I.Pr(peak) might be consistent with the research (Schwarzkopf et al., 1988, Viswanathan et al., 2008) which contended the applicability of top-down forecasting for highly variable data; however, the case with I.Cv(size) might be inconsistent with the research. However, it should be made aware that the result does not firmly corroborate or violate the research because the previous research employed rather different kinds of measures (Schwarzkopf et al., 1988) or a different level of forecasting (i.e. group level forecasting) (Viswanathan et al., 2008).

Forecasting horizon

The association between forecasting horizon and the performance of hierarchical forecasting in the literature (Shlifer and Wolff, 1979) was not repeated as PROLT was excluded from all of the models tested.

UV and DV

Among the statistics representing historical unit volume (UV) [G.Mean and I.Mean] and historical dollar volume (DV) [unit purchasing price × G.Mean and unit purchasing price × I.Mean], I.Mean was observed to be a significant predictor for the final model, and G.Mean was also observed to be a significant predictor for a model using forward stepwise method for the 10% cross-validation. By the e^b, the odds of classifying *SCtqum* increase with the increasing either G.Mean or I.Mean. These results might be consistent with the argument of Fliedner and Mabert (1992). They argued that UV is a significant grouping criterion which increases the accuracy of top-down forecasting.

Trend

Among the statistics [G.Slope and Slope] representing the linear trend, I.Slope was observed to be a significant predictor for the final model, and G.Slope was observed to be a significant predictor for some of the models using forward or backward stepwise method for the 10% cross-validation. By the e^b, the odds of classifying *SCtqum* increase with the increasing either I.Slope or G.Slope.

Deviation from a normal distribution

Among the statistics [G.Kurtosis, G.Skewness, I.Kurtosis and I.Skewness] representing the deviation from a normal distribution, I.Kurtosis was observed to be a significant predictor for the final model, and G.Kurtosis was observed to be a significant predictor for some of the models using forward or backward stepwise method for the 10% cross-validation. By the e^b, the odds of classifying *SCtqum* increase with the increasing either I.Kurtosis or G.Kurtosis.

Intermittency

As stated earlier, the intermittency was expected to be a significant demand feature which influences upon the performance of hierarchical forecasting. This is because the intermittency is an important demand feature which guides the use of forecasting methods for non-normal demand in the literature (Johnston and Boylan, 1996, Boylan et al., 2008). However, all of the statistics representing intermittency [G.Pr(zero) and I.Pr(zero)] were found to be non-significant for all of the models tested.

Categorical variables

Between the two categorical predictors, Equipment was observed to be a significant predictor. The odds of classification in Gun/RD (i.e. the reference group) were significantly different from the odds in ME (i.e. Equipment 1) and GE/AC (i.e. Equipment 2) by the Wald statistic. The odds of classification in the year 2005 (i.e. the reference group) were significantly different from the odds in the year 2007 (i.e. Year 2); however, they were non-significantly different from the odds in the year 2006 (i.e. Year 1) by the Wald statistic.

It is noteworthy that there were no group level demand features in the final model and only three group level demand features (i.e. G.Slope, G.Kurtosis and G.Mean) of the five group level demand features (as shown in Table 6-4) in the employed models for the 10% cross-validation found to be significant. As stated in Subsection 6.3.5, this might imply that demand features at group level have less effect on the relative performance of the alternative forecasting methods than demand features at item level.

So far, the associations between the forecasting methods and the multivariate demand features have been discussed within the logistic regression classification model. However, the individual associations between the forecasting methods and the demand features can be different. In the following subsection, the individual relations between the relative performance of the alternative forecasting methods and the significant predictors are analysed.

6.4.4 Individual relations between forecasting methods and demand features

The relative performance of the alternative forecasting methods is equivalent to the ideal selection assessed by equation (6-1): an observation (i.e. item i at time t) which shows smaller error by equation (6-1) for *tsm* (or *SCtqum*) than that for *SCtqum* (or *tsm*) was classified as an observation for *tsm* (or *SCtqum*). The associations between the individual variables which have b-coefficients significantly different from zero according to the Wald statistic in the models (either the final classification model or the employed classification models for the 10% cross-validation) and the relative performance of the alternative forecasting methods were investigated. For this analysis, all the 900 observations (i.e. the 300 items for the 3 years) used for the logistic regression model were examined. In order to examine the predictability of the demand features, the demand features in the previous period (not in the future period) were used in the same way as the classification model. "SPSS 15.0 for windows" was used to examine the individual relations.

As with the logistic regression classification models, the individual relations have been also examined between a combinatorial forecasting method (i.e. *SCtqum*) and a direct forecasting method (i.e. *tsm*), because they represented the most robust hierarchical and direct forecasting methods respectively. Therefore, it could not be asserted firmly that any result of this research is either consistent or inconsistent with previous research which compared top-down forecasting and direct forecasting.

Categorical variable

Table 6-15 provides cross-tabulation between the alternative forecasting methods and the three equipment groups. Expected count is defined as in equation (6-16).

$$Expected\ count = \frac{Row\ Total \times Column\ Total}{900} \tag{6-16}$$

Table 6-15 Cross-tabulation between the alternative forecasting methods and the three equipment groups

		tsm	*SCtqum*	Total
Gun/RD	Count	43	89	132
	Expected count	60.7	71.3	132.0
	% within equipment	32.6%	67.4%	100.0%
ME	Count	274	290	564
	Expected count	259.4	304.6	564.0
	% within equipment	48.6%	51.4%	100.0%
GE/AC	Count	97	107	204
	Expected count	93.8	110.2	204.0
	% within equipment	47.5%	52.5%	100.0%
Total	Count	414	486	900
	Expected count	414.0	486.0	900.0
	% within equipment	46.0%	54.0%	100.0%

In ME and GE/AC, more *tsm* and fewer *SCtqum* were observed than expected counts, whereas in Gun/RD fewer *tsm* and more *SCtqum* were observed than expected counts. This has an implication that *SCtqum* provided much better forecasts in Gun/RD than in other groups.

Pearson's χ^2 test can be used for testing whether there is a relationship between two categorical variables (Fisher, 1922). The test statistics can be written as in equation (6-17) (Kanji, 2006). The expected count should be greater than 5. Table 6-16 presents the result of Pearson's χ^2 test for independence between the relative performance of the alternative forecasting methods and the three equipment groups. The χ^2 value, 11.287, was significant as shown. This would indicate a significant association between the relative performance of the alternative forecasting methods and the three equipment groups. This corroborates the results from the logistic regression models.

$$\chi^2 = \sum \frac{(Observed\ count - Expected\ count)^2}{Expected\ count} \tag{6-17}$$

Table 6-16 Chi-Square test between the forecasts and the equipment groups

	Value	*df*	Significance
Pearson χ^2	11.287	2	0.004

The better performance of *SCtqum* for Gun/RD than that for the other two equipment groups can be explained by the relatively inferior performance of *tsm* to *um* for Gun/RD in terms of the mean rank for MAD and RMSE as shown in Subsection 5.4.3. For ME and GE/AC, *tsm* provided much better performance than that for Gun/RD. The differences in the relative performance between *SCtqum* and *tsm* for ME and GE/AC have been less obvious than the difference for Gun/RD because *tsm* was also a good forecasting method for ME and GE/AC. This might make the relative performance of *SCtqum* for Gun/RD look much better.

Continuous variables

As stated in Subsection 2.5.2, ANOVA was used to examine relations between grouping criteria and the performance of top-down forecasting by Fliedner and Mabert (1992) and between the number of groups and the performance of top-down forecasting by Fliedner and Lawrence (1995) with volatile monthly demand for spare parts. One way ANOVA involves one independent variable (Pallant, 2005). Table 6-17 presents the one way ANOVA results for the continuous variables which have *b*-coefficients significantly different from zero by looking at the Wald statistic in the classification models. The one independent variable was the relative performance (as defined earlier in this Subsection) of the alternative forecasting methods; the dependent variables were the variables. ANOVA tests the null hypothesis that all group means are equal.

Table 6-17 ANOVA for the variables in the logistic regression model

Variable		Sum of squares	*df*	Mean square	*F*	Sig.
G.Slope	Between Groups	819.237	1	189.237	0.645	0.422
	Within Groups	263277.3	898	293.182		
	Total	263466.6	899			
G.Kurtosis	Between Groups	1.019	1	1.019	0.51	0.821
	Within Groups	17862.834	898	19.892		
	Total	17863.853	899			
G.Mean	Between Groups	14347.817	1	14347.817	0.360	0.549
	Within Groups	35792162	898	39857.642		
	Total	35806510	899			
I.Slope	Between Groups	3.169	1	3.169	2.258	0.133
	Within Groups	1260.682	898	1.404		
	Total	1263.851	899			
I.Cv(size)	Between Groups	5.550	1	5.550	13.403	0.000
	Within Groups	371.883	898	.414		
	Total	377.434	899			
I.Corr(group)	Between Groups	.198	1	.198	2.941	0.087
	Within Groups	60.515	898	.067		
	Total	60.713	899			
I.Kurtosis	Between Groups	1036.833	1	1036.833	6.384	0.012
	Within Groups	145834.684	898	162.399		
	Total	146871.517	899			
I.Pr(peak)	Between Groups	0.021	1	0.21	0.906	0.341
	Within Groups	20.770	898	0.23		
	Total	20.790	899			
I.Mean	Between Groups	1.976	1	1.976	0.001	0.975
	Within Groups	1843382.347	898	2052.764		
	Total	1843384.324	899			

The statistics (i.e. variables), I.Cv(size) and I.Kurtosis had significant associations with the relative performance of the alternative forecasting methods as shown. The alternative hypothesis of associations between the two statistics [i.e. I.Cv(size) or I.Kurtosis] and the relative performance of the alternative forecasting methods should be accepted, with only a less than one in a thousand chance of making a type 1 error for I.Cv(size), and 1.2% chance of making a type 1 error for I.Kurtosis. However, the statistics, G.Slope, G.Mean, G.Kurtosis, I.Pr(peak), I.Slope, I.Corr(group) and I.Mean had no association with the relative forecasting performance as shown.

The reason why I.Slope, I.Corr(group) and I.Mean were significant predictors in the final model in spite of the non-significance for the individual relations might be attributed to the suppressor effect. Although there are no significant direct associations between the statistics [i.e. I.Slope, I.Corr(group) or I.Mean] and the relative forecasting performance in terms of one way ANOVA, these statistics contribute to the final logistic regression model with another predictor held constant. The significance of the predictors in the classification models for the 10% cross-validation sets might depend on the characteristics of data set. Although their individual associations were non-significant as shown in Table 6-17, G.Slope, G.Mean, G.Kurtosis and I.Pr(peak) were observed to be significant predictors in some of the classification models for the 10% cross-validation sets. This might be attributed to the different characteristics of different data sets making these predictors either significant or non-significant.

The ANOVA results for G.Slope, I.Slope, G.Kurtosis, I.Corr(group), G.Mean, I.Mean, and I.Pr(peak) reached different conclusions from the results of the logistic regression models which have presented significant associations. The ANOVA results for correlations were consistent with the previous literature (Dangerfield and Morris, 1992, Widiarta et al., 2008a, Widiarta et al., 2009) which claimed that correlations do not have any significant effect upon the relative forecasting performance between direct forecasting and top-down forecasting. The ANOVA results for G.Mean and I.Mean does not corroborate the research (Fliedner and Mabert, 1992) which has claimed that UV is a significant grouping criterion which increased the accuracy of hierarchical forecasting significantly.

Significant continuous variables

Table 6-18 and Table 6-19 describe the two variables which have *b*-coefficients significantly different from zero in terms of the Wald statistic as well as significant associations with the relative forecasting performance in terms of the ANOVA. As the e^b was 0.415 and 1.024 for I.Cv(size) and I.Kurtosis respectively in Table 6-13, the value of I.Cv(size) was expected to increase with a higher possibility of classifying *tsm*; whereas the value of I.Kurtosis was expected to increase with higher possibility of classifying *SCtqum*. The case with I.Cv(size) was confirmed in Table 6-18; that is, the mean level of *tsm* (i.e. 2.1925) was greater than the mean level of *SCtqum* (i.e. 2.0349). However, the case with I.Kurtosis was not confirmed in Table 6-19; that is, the mean level of *tsm* (i.e. 17.4279) was also greater than the mean level of *SCtqum* (i.e. 15.2762). The values of the both variables were observed to increase when the possibility of classifying *tsm* increases.

Table 6-18 Description of I.Cv(size)

	Mean	Std	SE	95% Confidence Interval for Mean		Min	Max
				Lower Bound	Upper Bound		
tsm	2.1925	.67993	.03342	2.1268	2.2582	.56	4.89
SCtqum	2.0349	.61082	.02771	1.9805	2.0893	.59	4.63
Total	2.1074	.64795	.02160	2.0650	2.1498	.56	4.89

Table 6-19 Description of I.Kurtosis

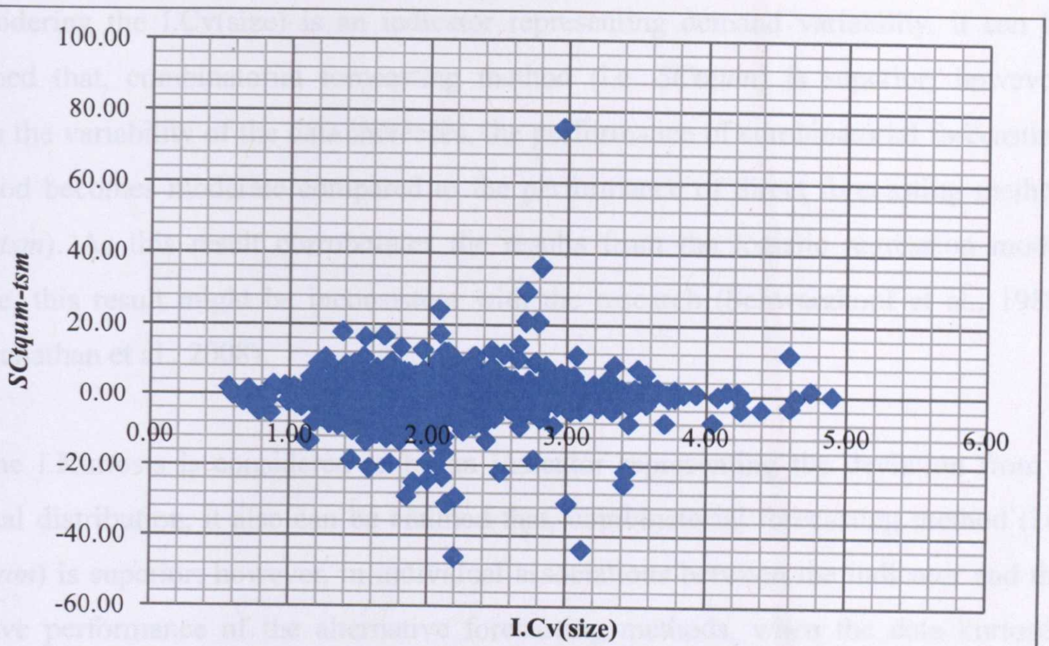
	Mean	Std	SE	95% Confidence Interval for Mean		Min	Max
				Lower Bound	Upper Bound		
tsm	17.4297	12.75752	.62700	16.1972	18.6622	-.08	56.06
SCtqum	15.2762	12.73174	.57752	14.1414	16.4109	-.38	58.37
Total	16.2668	12.78171	.42606	15.4306	17.1030	-.38	58.37

This inconsistency of I.Kurtosis could be explained by the correlations between I.Cv(size) and I.Kurtosis as shown in Table 6-20. The two variables strongly correlated in a positive direction as shown. These two statistics associate with the relative performance in the same direction, when the associations between the relative forecasting performance and the variables are detected individually. However, the two statistics were observed to influence the relative forecasting performance toward the different directions in the logistic regression model by the e^b.

Table 6-20 Correlation between I.Cv(size) and I.Kurtosis

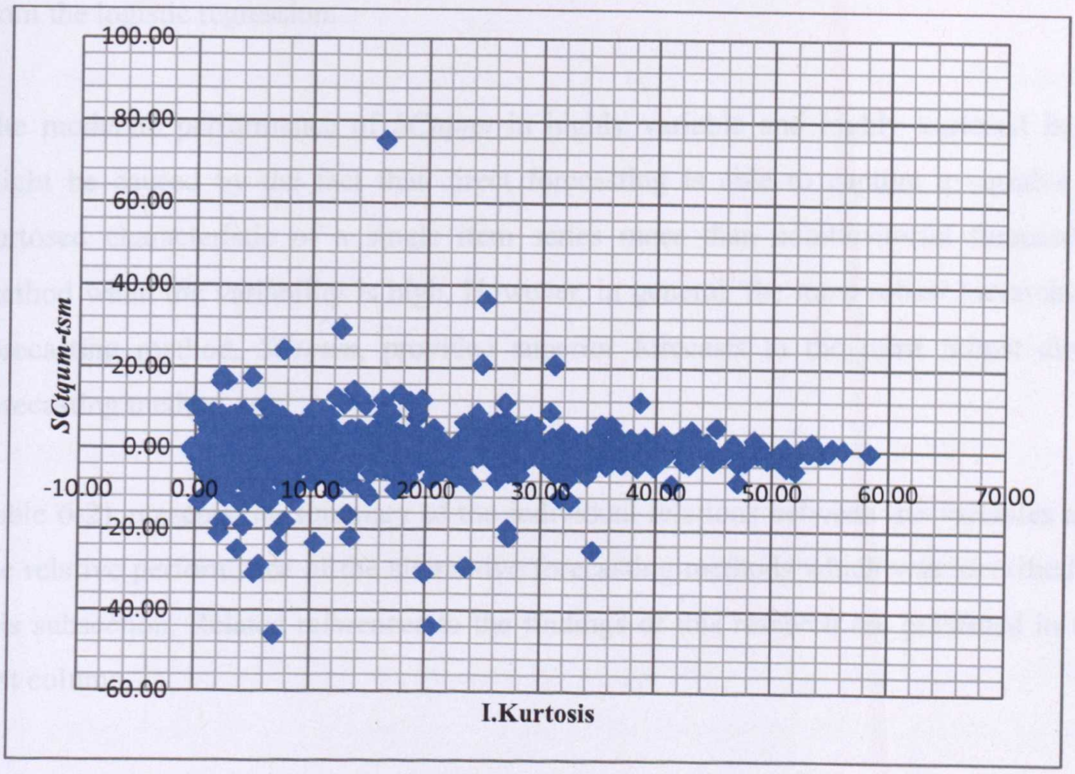
		I.Cv(size)	I.Kurtosis
I.Cv(size)	Pearson Correlation	1	0.732
	Sig. (2-tailed)		0.000
I.Kurtosis	Pearson Correlation	0.732	1
	Sig. (2-tailed)	0.000	

The individual relations for the 900 observations can be observed in Figure 6-6 and Figure 6-7. The value in the y-axis (i.e. *SCtqum-tsm*) smaller than zero indicates that *SCtqum* is superior to *tsm*; the value greater than zero in the y-axis indicates that *tsm* is superior to *SCtqum*. In most cases, the dots place around the zero point in y-axis. This means that the difference of the performance is indistinct. It was difficult to identify a cut-off value which divides the demand features clearly showing superior forecasting preferences. However, there was an obvious tendency that *SCtqum* was superior to *tsm* in lower value of I.Cv(size) and I.Kurtosis.



SCtqum-tsm = the absolute deviation divided by the monthly mean of *SCtqum* deducted by that of *tsm*.

Figure 6-6 Relations between *SCtqum-tsm* and I.Cv(size)



SCtqum-tsm = the absolute deviation divided by the monthly mean of *SCtqum* deducted by that of *tsm*.

Figure 6-7 Relations between *SCtqum-tsm* and I.Kurtosis

Considering the I.Cv(size) is an indicator representing demand variability, it can be claimed that, combinatorial forecasting method (i.e. *SCtqum*) is superior; however, when the variability of the data increases, the performance of combinatorial forecasting method becomes moderate compared to the performance of direct forecasting method (i.e. *tsm*). As this result corroborates the results from the logistic regression model above, this result might be inconsistent with the research (Schwarzkopf et al., 1988, Viswanathan et al., 2008).

As the I.Kurtosis is considered to be an indicator representing the deviation from a normal distribution, it also can be claimed that, combinatorial forecasting method (i.e. *SCtqum*) is superior; however, in individual associations between the indicator and the relative performance of the alternative forecasting methods, when the data kurtosed more from a normal distribution, the performance of combinatorial forecasting method becomes moderate compared to the performance of direct forecasting method (i.e. *tsm*). Nevertheless, it should be noted that the case of I.Kurtosis is inconsistent with the result from the logistic regression.

The moderate performance of *SCtqum* in highly variable and highly kurtosed items might be caused by the fact that direct forecasting is able to capture a variable or kurtosed characteristic of a single item series more than combinatorial forecasting method when the variability is high. However, in general, the most robust hierarchical forecasting method, *SCtqum*, provided superior forecasts to the most robust direct forecasting method, *tsm*.

Table 6-21 presents the summary of the individual relations between the variables and the relative performance of the alternative forecasting methods which was identified in this subsection. Related references to the findings of this research are presented in the last column.

Table 6-21 Individual relations between forecasting methods and variables

Demand feature	Variable	Relation	Reference
Correlation	I.Corr(group)	Non-sig.	= Dangerfield and Morris (1992), Widiarta et al. (2008a) & Widiarta et al. (2009) ≠ Schwarzkopf et al. (1988), Gross and Sohl (1990), Fliedner (1999) and Widiarta et al.(2006)
Variability	I.Cv(size) I.Pr(peak)	I.Cv(size) → *tsm* Non-sig.	≠ Schwarzkopf et al. (1988) & Viswanathan et al.(2008)
UV & DV	G.Mean I.Mean.	Non-sig.	≠ Fliedner and Mabert (1992) & Fliedner and Lawrence (1995)
Deviation from a normal distribution	I.Kurtosis G.Kurtosis	I.Kurtosis → *tsm*	-
Trend	I.Slope G.Slope	Non-sig.	-
Categorical variable	Equipment	*SCtqum* provides much better forecasts for Gun/RD than other groups	-

→: increasing the value of the variable increases the relative performance of the forecasting method (*SCtqum* or *tsm*); = (or ≠): consistent (or inconsistent) with the finding of the research; Non-sig.: non-significance; UV: historical unit volume; DV: historical dollar volume.

6.5 Summary and Conclusion

This chapter proposed a multivariate classification model to predict the relative performance of the alternative forecasting methods between the most robust hierarchical forecasting method (i.e. *SCtqum*) and the most robust direct forecasting method (i.e. *tsm*).

6.5.1 Summary of findings

For the purpose of predicting the relative performance of the alternative forecasting methods, a logistic regression model was built. Then, the performance of the logistic regression classification model was validated with separated 10% test data sets using the 10% cross-validation. In 7 test sets out of all 10 test sets, the classification model presented smaller forecasting errors than the result when using only the most robust forecasting method (i.e. *SCtqum*). In 6 test sets, the model presented smaller total inventory costs than the result using only *SCtqum*. The sum of the forecasting errors (the absolute deviations divided by monthly mean) reduced from 7,154.8 in the result using only *SCtqum* to 7,132.9 in the result from the 10 (10%) test sets using the classification model; the total inventory costs reduced from ₩693,601,747 (£354,420) in the result using only *SCtqum* to ₩679,816,280 (£347,377) in the result from the 10

(10%) test sets using the classification model as shown in Figure 6-4. While the total percentage of correctness in the result using only *SCtqum* was 54.0, the total percentage of correctness in the result from the 10 (10%) test sets using the classification model was 55.9% as shown Table 6-10. Hence, it might be suggested that the internal validation is established for the logistic regression classification model.

The final model was built using backward stepwise method. In the final model, among the 19 continuous variables and the 2 categorical variables, the 5 continuous variables [i.e. I.Slope, I.Cv(size), I.Corr(group), I.Kurtosis and I.Mean] and the 1 categorical variable (Equipment) were significantly different from 0. Additionally, 4 more predictors of the employed models were found to be significant for the 10% cross-validation; that is, G.Slope, G.Mean, G.Kurtosis and I.Pr(peak). No group level demand feature (i.e. predictors) of the final model and only 3 group level demand features of the employed models for the 10% cross-validation were found to be significant. Item level demand features might have more effect upon the performance of the classification model.

The individual relations between the relative performance of the alternative forecasting methods and the significant predictors in the logistic regression model were also analysed. Different results from the logistic regression model were observed. While G.Slope, G.Kurtosis, I.Slope, I.Corr(group), G.Mean, I.Pr(peak) and I.Mean were non-significant demand features in the ANOVA, the reason why these dead features were observed to be statistically significant predictors in the logistic regression models might be attributed to the suppressor effect in the logistic regression model and the different characteristics of data in different cross-validation data sets.

The summary of the findings in this chapter with respect to the previous research is presented as shown in Table 6-22. The summary of findings from the logistic regression model and the individual relations are presented in the left four and right two columns respectively. The findings need to be interpreted with caution. Dissimilar to the previous research which had compared top-down forecasting and direct forecasting, this research has compared combinatorial forecasting and direct forecasting. This was because a forecasting method using a combinatorial forecasting strategy (i.e. *SCtqum*) was found to be robust as with the research (Kahn, 1998, Dekker et al., 2004, Hyndman et al., 2007). However, the previous research is worth comparing to the results of this research because combinatorial forecasting can be considered to be a variant of top-down forecasting. This is also a contribution of this research in that this research has investigated the influence of demand features upon the performance of a combinatorial forecasting method. There are more contributions in this research in that this research identified a combined influence of demand features of non-normal data including correlations and intermittency upon the performance of hierarchical forecasting.

Table 6-22 The influence of demand features upon the performance of the alternative forecasting methods

		Logistic regression classification model		Individual relations	
Demand feature	Predictor (variable)	Odds of classification	Reference	Relation	Reference
Correlation	I.Corr(group)	I.Corr(group) → *tsm*	= Schwarzkopf et al. (1988); ≠ Gross and Sohl (1990), Dangerfield and Morris (1992), Fliedner (1999), Widiarta et al.(2006), Widiarta et al. (2008a), & Widiarta et al. (2009)	Non-sig.	= Dangerfield and Morris (1992), Widiarta et al. (2008a) & Widiarta et al. (2009) ≠ Schwarzkopf et al. (1988), Gross and Sohl (1990), Fliedner (1999) & Widiarta et al. (2006)
Variability	I.Cv(size) I.Pr(peak)	I.Cv(size) (or I.Pr(peak)) → *tsm* (or *SCtqum*)	≠ (or =) Schwarzkopf et al. (1988) & Viswanathan et al. (2008)	I.Cv(size) → *tsm*	≠ Schwarzkopf et al. (1988) & Viswanathan et al. (2008)
Forecasting horizon	Non-sig.	-	≠ Shlifer and Wolff (1979)	-	-
UV & DV	G.Mean I.Mean	G.Mean or I.Mean → *SCtqum*	= Fliedner and Mabert (1992) & Fliedner and Lawrence (1995)	Non-sig.	≠ Fliedner and Mabert (1992)
Trend	G.Slope I.Slope	G.Slope or I.Slope → *SCtqum*	-	Non-sig.	-
Deviation from a normal distribution	G.Kurtosis I.Kurtosis	G.Kurtosis or I.Kurtosis → *SCtqum*	-	I.Kurtosis → *tsm*	-
Intermittency	Non-sig.	-	-	-	-
Categorical variable	Equipment	The odds of classification in Gun/RD are significantly different from the odds in ME and GE/AC	-	*SCtqum* provides much better forecasts for Gun/RD than other groups	-

→: increasing the value of the predictor (variable) increases the odds of classifying the forecasting method (or the relative performance of the forecasting method);
= (or ≠): consistent (or inconsistent) with the finding of the research; Non-sig.: non-significance; UV: historical unit volume; DV: historical dollar volume.

6.5.2 Forecasting scheme for the South Korean Navy

It was demonstrated that the most robust forecasting method for the spare parts demand is *SCtqum* in Chapter 5. Based on this finding, the forecasting scheme was suggested in Subsection 5.6.2. In particular, an attempt to select a robust forecasting method for spare parts with different demand features in ME was made. However, this forecasting scheme could not explain the effect of the demand features upon the relative forecasting performance explicitly, so that it might be difficult to apply to different data sets which have not been tested.

In this chapter, the process of building the logistic regression classification model was clarified. The logistic regression classification model might be able to be applied to different data sets because this model is based on the relationships between demand features and relative forecasting performance. This can be more generalisable than the forecasting scheme based on the different demand features in ME as suggested in Subsection 5.6.2. A forecasting scheme for the South Korean Navy based on demand features can be suggested as follows:

a) The logistic regression classification model provides a clear guideline to choose a superior forecasting method. It is recommended to implement the logistic regression classification model to select a forecasting method from the alternative forecasting methods (*tsm* vs. *SCtqum*) for forecasting spare parts demand.
b) Continuous demand features such as correlations [expressed as I.Corr(group)], variability [expressed as I.Cv(size) and I.Pr(peak)], trend [expressed as G.Slope and I.Slope], deviation from a normal distribution [expressed as G.Kurtosis and I.Kurtosis], and historical unit volume (UV) [expressed as G.Mean and I.Mean], and a categorical variable such as the type of equipment [expressed as Equipment] should be considered as predictors for the logistic regression classification model.
c) When selecting a predictor entry method for the logistic regression model, the two step process which chooses a superior stepwise method for the model is recommended. At the first step, forward stepwise method is preferentially considered. At the second step, if forward stepwise method selects too small a number of predictors (i.e. two predictors), backward stepwise method is suggested to be employed. Otherwise, forward stepwise method is suggested to be employed.

6.5.3 Conclusion

This chapter answers research question b) "what forecasting method is appropriate for the spare parts demand in the South Korean Navy?" by identifying the superior forecasting method in given demand features with the logistic regression classification model; answers research question c) "under what conditions are top-down forecasting or combinatorial forecasting superior or inferior to direct forecasting?" by identifying significant predictors and their coefficient in the logistic regression model as well as the individual relations between the demand features and the relative forecasting performance; and answers research question d) "how can the spare parts demand be classified in order to predict a superior forecasting method?" by proposing the multivariate logistic regression classification model.

In the next chapter, the findings of this research are summarised. Then, contributions, implications, limitations and further research are presented.

Chapter 7. Conclusion

The aim of this research was "to establish an appropriate forecasting strategy for predicting the demand for spare parts in the South Korean Navy". An appropriate forecasting strategy has been identified by achieving the research objectives; and the research objectives were achieved by answering the research questions. This chapter restates the research objectives and the research questions; then summarises the findings related to the research objectives and the research questions in Sections 7.1, 7.2 and 7.3. The contributions of this research are described in Section 7.4. A forecasting scheme for the South Korean Navy is suggested in Section 7.5. Limitations of this research are presented in Section 7.6. Finally, further research is recommended in Section 7.7.

7.1 Nature of the Spare Parts Demand

In order to achieve the research objective a) "to clarify the nature of the spare parts demand in the South Korean Navy", research question a) "what is the nature of the spare parts demand in the South Korean Navy?" was answered in Chapter 4. For answering research question a), 300 spare parts were selected from 9,369 spare parts (for the 3 types of warships in the South Korean Navy with time boundary from January 2002 and November 2007) and analysed. The demand for the 300 spare parts could account for approximately 60% of demand for the 9,369 spare parts

The time series of the spare parts tested were found to be non-normal; however, they correlated within a pair group. The non-normality of spare parts demand in militaries found in the literature (Markland, 1970, Businger and Read, 1999, Eaves and Kingsman, 2004) was repeated in this research.

Some relative demand features which are different in each equipment group were identified. Gun/RD (consisting of Gun I, II, III, and Radar I) was characterised as having higher intermittency, smaller demand volume, shorter lead time, and more expensive prices. ME (consisting of Main Engine I and II) was characterised as having lower correlation, steeper downward trend, lower intermittency, and larger demand volume. GE/AC (consisting of Generator I and Air Compressor I) was characterised as having higher variability, greater peakedness, and greater deviation from a normal distribution.

Features which could make hierarchical forecasting more accurate than direct forecasting for the 300 Naval spare parts demand were identified as:

a) Long forecasting horizons ranging from 15 months to 30 months for the Naval spare parts consisted of procurement lead time and review cycle imply a feature that hierarchical forecasting can be superior to direct forecasting (Shlifer and Wolff, 1979).
b) Less variable and less intermittent, and less peaked demand features at group level than those at item level [as Cv(size), Pr(zero), Pr(peak) at group level time series were smaller than those at item level time series] were found. Therefore, reduced non-normal demand features at group level time series can be characterised, so as to satisfy the major premise of hierarchical forecasting which could be superior to direct forecasting (Gross and Sohl, 1990, Fliedner and Lawrence, 1995, Fliedner, 2001).
c) The data obtained from the Navy were identified to contain missing or unreliable data. Hierarchical forecasting could present better performance than direct forecasting for predicting demand based on such data (Schwarzkopf et al., 1988).
d) The substitutability of the Naval spare parts is also a feature which could make hierarchical forecasting more accurate than direct forecasting. Hierarchical forecasting uses the historical demand at group level, which could be less dependent upon the degree of item substitutability than the direct demand at item level (Widiarta et al., 2008b).
e) The Naval spare parts are structured by the types of equipment and the National Stock Number (NSN) code. These represent the hierarchical structure of spare parts. Hierarchical forecasting is an advantageous forecasting strategy for forecasting demand which is structured by a hierarchical demand structure (Hyndman et al., 2007).

However, as the current forecasting methods in the South Korean Navy are too naïve to catch the characteristics of the Naval spare parts demand, they are inappropriate for the nature of the demand.

7.2 The Performance of the Forecasting Methods

In order to achieve the research objective b) "to compare the performance of the alternative forecasting strategies (i.e. top-down forecasting, combinatorial forecasting and direct forecasting) for predicting the spare parts demand at item level under the inventory control environment of the South Korean Navy", research questions b) "what forecasting method is appropriate for the spare parts demand in the South Korean Navy?" and c) "under what conditions are top-down forecasting or combinatorial forecasting superior or inferior to direct forecasting?" were answered in Chapter 5. For answering research questions b) and c), a range of forecasts using alternative forecasting methods were generated with the time series of the 300 items, and these were compared in terms of absolute, relative and derivative measures.

Derivative measures use simulation to derive the impact of forecasting accuracy in terms of the inventory levels and the service levels achieved by the inventory system. Two approaches (i.e. safety margin approach vs. total inventory costs approach) toward the total inventory costs in the simulation were compared. Between them, the total inventory costs approach was employed in this research. This is because it is unrealistic to adjust a safety stock to the exact amount of no stock-out as the case with safety margin approach which was used by some researchers (Wemmerlöv, 1989, Eaves, 2002, Eaves and Kingsman, 2004).

Four years (2004 ~ 2007) of yearly forecasts for the Naval spare parts were generated. Among them, the performance of the forecasts in 2004 was different from that of the forecasts in 2005 ~ 2007. This might be attributed to the influence of the two peak points in 2002 and 2003 upon the forecasts in 2004. A summary of findings about the performance of forecasting methods in 2005 ~ 2007 is presented as in Table 7-1. The most robust direct forecasting method among direct forecasting methods tested with the exception of the total inventory costs was found to be the simple exponential smoothing model with monthly aggregated data adjusted for linear trend and additive seasonality (*tsm*). In general, the most robust forecasting method among all the forecasting methods tested was found to be the simple combination between the simple exponential smoothing model with quarterly aggregated data adjusted for linear trend at group level and the simple exponential smoothing model with monthly aggregated unadjusted data at item level (*SCtqum*). Simple combination dominated Top-21 and Top-20. These are consistent with the literature (Kahn, 1998, Dekker et al., 2004, Hyndman et al., 2007) in that combinatorial forecasting could present lower forecasting errors and lower inventory costs than top-down and direct forecasting; and DeLurgio (1998) in that simple combination could be as good as a more complex proration method (e.g. weighted combination).

Table 7-1 A summary of findings: the performance of forecasting methods

		Absolute & relative measures	Derivative measure
DF	The most robust DF	*tsm*	Total inventory costs: *um* Mean rank for total inventory costs: *tsm*
DF	The most robust DF for equipment group	Gun/RD: *um*; ME & GE/AC: *tsm*	
HF	The most robust forecasting method	*SCtqum*	
HF	Proration method	Top-21 by LN(ratio): 14 SCs, 6 WCs, 0 TD1, & 1 TD2 Top-21 by mean rank in MAD/RMSE: 15 SCs, 5 WCs, 0 TD1, & 1 TD2	Top-20 by total inventory costs: 12 SCs, 8 WCs, 0 TD1, & 0 TD2
HF	The most frequently higher ranked DF used for Top-21 or Top-20	Item level: m^1 & u^2 Item level combination: *um* Group level: m^1 & q^1; t^2& ts^2 Group level combination: *tq*	Item level: m^1& u^2 Item level combination: *um* Group level: m^1 & t^2
HF	The most robust forecasting method for equipment group	ME: *SCtquy*	

DF = direct forecasting; HF = hierarchical forecasting; SC = simple combination; WC = weighted combination; TD = top-down forecasting; LN(ratio) = natural log relative error [ln($error_{HF}$/$error_{DF}$)]; *1* = data aggregation method; *2* = data adjustment method.

The most frequently higher ranked direct forecasting methods for the top 20 or 21 hierarchical forecasting methods are summarised in the table. As a combination of data aggregation method and data adjustment method, *um* was found to be the most frequently higher ranked method for the top 20 and 21 hierarchical forecasting methods at item level. The most robust forecasting methods for Main Engines were also identified as shown. This suggested a forecasting scheme for the South Korean Navy with respect to the different demand feature in ME. 3.8 % of the total inventory costs, which resulted from using the generally most robust forecasting method (i.e. *SCtqum*), were demonstrated to be reduced by using this forecasting scheme.

7.3 Forecasting Performance and Demand Features

For the purpose of achieving the research objectives c) "to investigate the influence of demand features upon the performance of the alternative forecasting strategies" and d) "to develop a classification model for the spare parts demand in order to predict a superior forecasting method", research questions c) "under what conditions are top-down forecasting or combinatorial forecasting superior or inferior to direct forecasting?" and d) "how can the spare parts demand be classified in order to predict a superior forecasting method?" were answered in Chapter 6. For answering research questions c) and d), the logistic regression classification model was implemented, and the individual relations between the relative performance of the alternative forecasting methods and the significant predictors in the logistic regression classification model were analysed.

The logistic regression classification model could predict a superior forecasting method between the two alternative forecasting methods (i.e. *SCtqum* and *tsm*). The sum of absolute deviations divided by the monthly mean demand (or the total inventory costs) over 900 observations (i.e. the 300 spare parts × 3 times of yearly forecasts) reduced from 7,154.8 (or £354,420) in the results using only *SCtqum* to 7,132.9 (or £347,377) in the result using the logistic regression classification model for the 10 (10%) test sets. While the total percentage of correctness using only *SCtqum* was 54.0, the total percentage of correctness using the logistic regression classification model for the 10 (10%) test sets was 55.9%. Hence, it might be suggested that the internal validation is established for the logistic regression classification model.

For the final model which uses all the 900 observations, the 5 continuous variables which represent trend [I.Slope], variability [I.Cv(size)], correlation [I.Corr(group)], deviation from a normal distribution [I.Kurtosis], and historical unit volume [I.Mean] and the 1 categorical variable [Equipment] were observed to be statistically significant. For the logistic regression classification models for the 10% cross-validation which used 90% of the observations (i.e. 10 training sets), 4 more continuous variables which represent trend [G.Slope], variability [I.Pr(peak)], deviation from a normal distribution [G.Kurtosis] and historical unit volume [G.Mean] were observed to be statistically significant. For the final model, all of the significant predictors were found to be item level demand features; none of them were found to be group level demand features. For the logistic regression classification models for the 10% cross-validation, only 3 significant group level predictors were observed. Item level demand features might have more effect upon selecting a superior forecasting method in the logistic regression model. A summary of findings about the influence of demand features upon the relative performance of the alternative forecasting methods with respect to the related previous research in terms of the logistic regression model and the individual relations is presented as shown in Table 7-2.

The influence of the significant demand features in the logistic regression model was identified. As G.Slope, I.Slope, G.Kurtosis, I.Kurtosis, G.Mean, I.Mean, or I.Pr(peak) increases, the odds of classifying *SCtqum* increase. On the contrary, as I.Cv(size) or I.Corr(group) increases, the odds of classifying *tsm* increase. The case with I.Corr(group) might be consistent with Schwarzkopf et al. (1988), the case with I.Pr(peak) might be consistent with the research (Schwarzkopf et al., 1988, Viswanathan et al., 2008); the cases with G.Mean and I.Mean might be consistent with Fliedner and Mabert (1992).

The individual relations between the relative performance of the alternative forecasting methods and the significant predictors in the classification model were more or less inconsistent with the results from the logistic regression classification model. *SCtqum* provided much better forecasts for Gun/RD than for ME and GE/AC. G.Slope, I.Slope, I.Corr(group), G.Mean, I.Mean, G.Kurtosis, and I.Pr(peak) were found to be non-significant demand features which are inconsistent with the logistic regression classification model results. I.Cv(size) and I.Kurtosis were found to increase with higher possibility of classifying *tsm*. The result with I.Kurtosis is inconsistent with the logistic regression classification model result. The non-significance of I.Corr(group) might be consistent with the research (Dangerfield and Morris, 1992, Widiarta et al., 2008a, Widiarta et al., 2009).

Table 7-2 A summary of findings: the influence of demand features upon the relative performance of alternative forecasting methods

	Logistic regression classification model		Individual relations	
Demand feature	Odds of classification	Reference	Relation	Reference
Correlation	I.Corr(group) → *tsm*	= Schwarzkopf et al. (1988); ≠ Gross and Sohl (1990), Dangerfield and Morris (1992), Fliedner (1999), Widiarta et al.(2006), Widiarta et al. (2008a), & Widiarta et al. (2009)	Non-significance	= Dangerfield and Morris (1992), Widiarta et al. (2008a) & Widiarta et al. (2009) ≠ Schwarzkopf et al. (1988), Gross and Sohl (1990), Fliedner (1999) & Widiarta et al. (2006)
Variability	I.Cv(size) (or I.Pr(peak)) → *tsm* (or *SCtqum*)	≠ (or =) Schwarzkopf et al. (1988) & Viswanathan et al. (2008)	I.Cv(size) → *tsm*	≠ Schwarzkopf et al. (1988) & Viswanathan et al. (2008)
Forecasting horizon	Non-significance	≠ Shlifer and Wolff (1979)	-	-
UV & DV	G.Mean or I.Mean → *SCtqum*	= Fliedner and Mabert (1992) & Fliedner and Lawrence (1995)	Non-sig.	≠ Fliedner and Mabert (1992)
Trend	G.Slope or I.Slope → *SCtqum*	-	Non-significance	
Deviation from a normal distribution	G.Kurtosis or I.Kurtosis → *SCtqum*	-	I.Kurtosis → *tsm*	-
Intermittency	Non-significance	-	-	-
Categorical variable	The odds of classification in Gun/RD are significantly different from the odds in ME and GE/AC		*SCtqum* provides much better forecasts for Gun/RD than other groups	

→: increasing the value of the predictor (variable) increases the odds of classifying the forecasting method (or the relative performance of the forecasting method); = (or ≠): consistent (or inconsistent) with the finding of the research; UV: historical unit volume; DV: historical dollar volume.

7.4 Contributions

Theories and guidelines for using hierarchical forecasting for predicting non-normal demand associated with the spare parts demand in the South Korean Navy are not well-developed. The research gaps led to the current research objectives. In the process of achieving these research objectives, the research contributions have been made. The contributions of this research are presented in Table 7-3.

In spite of the applicability of hierarchical forecasting in predicting volatile and intermittent demand as stated in Subsection 1.2.4, the literature has paid little attention to the use of hierarchical forecasting for the intermittent demand at item level. This is a feature of non-normal demand associated with spare parts demand. This research gap led to the first two research objectives. In the process of achieving the two research objectives, the nature of the spare parts demand and the performance of the alternative forecasting strategies were identified.

There have been no controlled studies on: the influence of correlations between non-normal demand time series upon the performance of the alternative forecasting strategies; the influence of intermittency upon the performance of the alternative forecasting strategies; the influence of demand features upon the performance of combinatorial forecasting; and, the combined influence of demand features of empirical non-normal data upon the performance of the alternative forecasting strategies. These research gaps led to the next two research objectives. In the process of achieving these two research objectives, a new multivariate logistic regression classification model for predicting the relative performance of the alternative forecasting methods was developed. In doing this the influence of the demand features upon the relative performance of the alternative forecasting methods was identified.

The forecasting performance was evaluated with the three groups of measurement. With these three-fold measurements, reliability and internal validity of the results were established. The practical impact of forecasting methods on the inventory system was evaluated with the derivative measures.

Table 7-3 Contributions of the research

Research gaps	Research objectives	Contributions
• Little attention to the use of hierarchical forecasting for the intermittent demand at item level.	• To clarify the nature of the spare parts demand in the South Korean Navy.	• The nature of the spare parts demand in the South Korean Navy was identified and this showed that it is non-normal and correlates with each other in a group. • The features of the spare parts demand, which could make hierarchical forecasting more accurate than direct forecasting, were identified, and these were: a reduction of non-normality at group level, long forecasting horizons, substitutability, and the unreliability of data.
	• To compare the performance of the alternative forecasting strategies for predicting the spare parts demand at item level under the inventory control environment of the South Korean Navy.	• The performance of the alternative forecasting strategies (hierarchical forecasting and direct forecasting) for the spare parts demand was demonstrated as shown in Table 7-1. • The robust forecasting methods which provided consistently superior forecasting performance for spare parts in each equipment group were also identified.
• Little research on the influence of correlations and intermittency of non-normal demands upon the relative performance of the alternative forecasting strategies (hierarchical forecasting and direct forecasting); • Little research on the influence of demand features upon the performance of combinatorial forecasting; • Little research on the combined influence of demand features of empirical non-normal data upon the relative performance of the alternative forecasting strategies.	• To investigate the influence of demand features upon the performance of the alternative forecasting strategies; • To develop a classification model for the spare parts demand in order to predict a superior forecasting method.	• A new multivariate logistic regression classification model for the spare parts demand, which predicts the relative performance of the alternative forecasting methods (*tsm* vs. *SCtqum*) by the combinations of the multivariate demand features {i.e. trend [G.Slope and I.Slope], variability [I.Cv(size) and I.Pr(peak)], correlation [I.Corr(group)], deviation from a normal distribution [G.Kurtosis and I.Kurtosis], historical unit volume [G.Mean and I.Mean], and Equipment}, was developed. • Intermittency was found to be non-significant variable in the logistic regression classification model. • The combined influences of the demand features of non-normal demand associated with spare parts demand in the logistic regression classification model upon the relative performance of the alternative forecasting methods were identified as shown in Table 7-2. • The logistic regression classification model can provide a clear guideline to select a superior forecasting method between *SCtqum* and *tsm*. • The improvement of the performance in the classification model compared to the performance using only the most robust forecasting method (i.e. *SCtqum*) was demonstrated. • The properties of the predictor entry methods were identified so that the two step process, which selects a superior stepwise method for the logistic regression classification model, was proposed. • The individual influences of the three demand features {i.e. a categorical variable [Equipment], variability [I.Cv(size)] and deviation from a normal distribution [I.Kurtosis]} of the data used for generating the forecasts upon the performance of the alternative forecasting methods (*tsm* vs. *SCtqum*), which predict the relative performance of the alternative forecasting methods, were identified.
		• In addition to absolute and relative measures, derivative measures using simulation were employed so as to maintain reliability and internal validity of the results. With simulation experiments using the empirical 300 spare parts data obtained from the South Korean Navy, the performance of the alternative forecasting strategies, the logistic regression classification model, and a forecasting scheme constructed for the South Korean Navy were verified. For the classification model, a range of diagnostics and the cross-validation as well as the simulation experiment established the internal validity.

7.5 Forecasting Scheme for the South Korean Navy

A forecasting scheme for predicting the demand for spare parts in the South Korean Navy is suggested in this research. The performance of this forecasting scheme was demonstrated with the three-fold accuracy measures. This forecasting scheme is presented as follows:

a) In lieu of the current direct forecasting, hierarchical forecasting is suggested.

b) As a proration method for hierarchical forecasting, combinatorial forecasting, especially simple combination, should be considered.

c) A careful selection of a forecasting method from various forecasting methods using simple combination is required, because the performance of the forecasting methods using simple combination for forecasting the 300 spare parts demand was highly variable.

d) As a forecasting method using simple combination, *SCtqum* is recommended for forecasting spare parts demand for Gun/RD and GE/AC, because *SCtqum* generally provided the most robust forecasting performance for forecasting the 300 spare parts demand.

e) For forecasting spare parts demand for ME which is characterised as having lower correlation, steeper downward trend, lower intermittency, and larger demand volume, *SCtquy* is recommended, because *SCtquy* provided the most robust forecasting performance for forecasting the spare parts demand for ME.

f) However, *SCtqum* and *SCtquy* might be difficult to use for the spare parts of other pieces of equipment or other types of warships. The logistic regression classification model can be used even for other demands, because this model is based on the relationships between demand features and the relative forecasting performance of the alternative forecasting methods.

g) The logistic regression classification model provides a clear guideline to choose a superior forecasting method. It is recommended to implement the logistic regression classification model to select a forecasting method from the alternative forecasting methods (*tsm* vs. *SCtqum*) for forecasting the spare parts demand.

h) Continuous demand features such as correlation [expressed as I.Corr(group)], variability [expressed as I.Cv(size) and I.Pr(peak)], trend [expressed as G.Slope and I.Slope], deviation from a normal distribution [expressed as G.Kurtosis and

I.Kurtosis], historical unit volume (UV) [expressed as G.Mean and I.Mean], and a categorical variable such as the type of equipment [expressed as Equipment] should be considered as predictors for the logistic regression classification model.

i) When selecting a predictor entry method for the logistic regression model, the two step process which chooses a superior stepwise method for the model is recommended. At the first step, forward stepwise method is preferentially considered. At the second step, if forward stepwise method selects too small a number of predictors (i.e. two predictors), backward stepwise method is suggested to be employed. Otherwise, forward stepwise method is suggested to be employed.

j) Verification of a forecasting performance using simulation before using the forecast for a procurement decision should be conducted. The simulation can reduce the risk of a wrong decision and guarantee the best practical decision in terms of inventory and service levels.

7.6 Limitations

Three important limitations need to be considered. The first limitation is the difficulty of obtaining and publishing detailed information. The Military might be one of the strictest organisations regarding issues of confidentiality. It is against the Military information Security Act to collect and reveal confidential information. Therefore, accessing detailed information about the spare parts demand and the inventory systems was limited. However, as the time series of the 9,369 spare parts were collected, there might be no problem to investigate an appropriate forecasting strategy for the spare parts.

The second limitation is the difficulty of obtaining true demand. The five sources of non-normality of the spare parts demand were identified in Section 4.5. Among them, the three sources (i.e. the multi-echelon inventory systems, the budgeting process and the maintenance system) would generate proxy demand data, because they distort true demand. Forecasts based on the proxy demand are likely to be biased. However, since no other data sources are available except the logistical database of the Naval Logistics Command, this research relies on the logistical database, and assumes that the construct validity is established.

The third limitation lies in the fact that the logistic regression classification model and the individual relations between the demand features and the performance of the alternative forecasting methods are based on the comparisons between combinatorial forecasting and direct forecasting rather than top-down forecasting and direct forecasting. As the previous research compared top-down forecasting and direct forecasting, it could not be asserted firmly that any result of this research on classification is either consistent or inconsistent with the previous research. As stated, this is a limitation as well as a contribution of this research in that this research has investigated the influence of demand features upon the performance of a combinatorial forecasting method which might not have been conducted before. The previous research is worth comparing with the result of this research because combinatorial forecasting can be considered to be a variant of top-down forecasting. As such, existing theories could provide frameworks for the findings of this research.

7.7 Further Research

This research has thrown up four questions in need of further investigation. First, in order to increase the generalisability of the classification model, more investigation with more data sets, either in other militaries or in other organisations might be required.

Second, it would be interesting to examine more demand features for the classification model as well as their individual influence upon the forecasting performance. For example, the influence of seasonality needs to be examined. Seasonality was not included in the logistic regression classification model in this research. This was because most of the seasonal effects were non-significant as shown in Subsection 4.3.5. However, some of the seasonal effects were found to be significant as shown in Subsection 4.3.5. *tsm* which is a forecasting method using data adjusted for trend and seasonality was observed to present the most robust performance among the direct forecasting methods tested as shown in Section 5.4. There could be interesting relations between seasonality and forecasting performance.

Third, further research should be done to investigate the impact of data adjustment methods on the outcomes of hierarchical forecasting methods. As shown in Table 7-1, the forecast with monthly aggregated data adjusted for linear trend and additive seasonality (*tsm*) was the most robust direct forecasting method; the most frequently higher ranked direct forecasting methods for the top 21 or the top 20 hierarchical forecasting methods were also identified. However, the impact of trend and/or seasonality adjustment on the performance of the 220 hierarchical forecasting methods might be different.

Fouth, in this research, the dichotomy classification and relations between one of the combinatorial forecasting methods (i.e. *SCtqum*) and one of the direct forecasting methods (i.e. *tsm*) were investigated. However, in some cases, other forecasting methods were observed to be higher ranked than *SCtqum* and *tsm*. For instance, *um* was observed to be higher ranked than *tsm* in terms of the total inventory costs in Subsection 5.4.4; and *TD2tsm* was observed to be higher ranked than *SCtqum* in terms of the natural log relative error [$\ln(\text{error}_{HF}/\text{error}_{tsm})$] for the total inventory costs in Subsection 5.5.3. Therefore, investigation employing other forecasting methods (e.g. *TD2tsm* or *um*) is needed.

Appendices

Appendix A - Probability Distributions

A probability distribution refers to a specification of all the values and their associated probabilities for a complete description of a random variable (Krzanowski, 1998). A variable whose numerical values are determined by chance factors is named a random variable (Krishnamoorthy, 2006). Random variables can be divided into discrete and continuous random variables. When the set of all possible values of a random variable X is countable, X is named a discrete random variable; and when the set of all possible values of X is an interval or union of two or more non-overlapping intervals, X is named a continuous random variable (Krishnamoorthy, 2006). Probability of an event, $P(A)$ can be expressed as in equation (A-1) (Krishnamoorthy, 2006).

$$P(A) = \frac{Number\ of\ outcomes\ in\ the\ event\ A}{Total\ number\ of\ outcomes\ in\ the\ sample\ space} \tag{A-1}$$

If R is the set of all possible values of a discrete random variable X, and $f(k) = P(X = k)$ for each k in R, then $f(k)$ is called the probability mass function of X where: $P(X = k)$ is the probability that X assumes the value k (Krishnamoorthy, 2006). For a continuous random variable any real valued function $f(x)$ which satisfies equation (A-2) is defined as a probability density function. Table A-1 describes various probability distributions and presents the equations of them.

$$f(x) \geq 0\ \ for\ all\ x,\ \ and\ \int_{-\infty}^{\infty} f(x)dx = 1 \tag{A-2}$$

Table A-1 Probability distributions (Krzanowski, 1998, Balakrishnan and Nevzorov, 2003, Krishnamoorthy, 2006)

Distribution	Description	Equation
Normal (Gaussian) distribution	The normal distribution is the most commonly used distribution. Let X has a normal distribution with mean μ and variance σ^2, and then the probability density function of the normal distribution can be expressed as in equation (A-3). The notation $X \sim N(\mu, \sigma^2)$ is used for indicating a variable with the normal distribution.	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad (-\infty < x < \infty)$ (A-3)
Uniform (Rectangular) distribution	The uniform (or rectangular) distribution is useful for modelling random phenomena related to continuous measurements such as time, length and area. Let X is a variable ranging $a \le x \le b$, the probability density function of the uniform distribution of X can be expressed as in equation (A-4). The notation $X \sim U(a, b)$ is used for indicating a variable with the uniform distribution.	$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise;} \end{cases}$ (A-4) where: $\mu = (a + b)/2$; $\sigma^2 = (b - a)^2/12$
χ^2 distribution	The χ^2 distribution is derived from the normal distribution. Let $X_1, X_2, X_3, \ldots, X_k$ are mutually independent $N(0, 1)$ variables, the χ^2 distribution on k degrees of freedom can be expressed as in equation (A-5).	$\chi^2 = \sum_{i=1}^{k} X_i^2$ (A-5)
F-distribution	The F-distribution is also derived from the normal distribution. Let χ_k^2 and χ_m^2 are independent, F-distribution on k and m degrees of freedom, $F_{k,m}$, can be expressed as in equation (A-6).	$F_{k,m} = \frac{\chi_k^2}{k} \div \frac{\chi_m^2}{m}$ (A-6)
Lognormal distribution	A positive random variable X is lognormally distributed, if $\ln(X)$ is normally distributed. The probability density function of the lognormal distribution of X can be expressed as in equation (A-7). The lognormal distribution can be used when the random variable X assumes only positive values and its histogram is skewed to the right.	$f(x) = \frac{b}{\sqrt{2\pi}x} e^{-\left\{\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}}$ (A-7)

(Continued)

Table A-1 Continued

Distribution	Description	Equation
Gamma distribution	The gamma distribution presents extremely flexible modelling possibilities. The gamma distribution has two parameters (a scale parameter, a, and a shape parameter, b) which can generate a wide selection of shapes of density curves. The probability density function of the gamma distribution can be written as in equation (A-8). A gamma distribution with a positive integer shape parameter, b, is called the Erlang distribution. The gamma function, $\Gamma(x)$ is defined as in equation (A-9).	$f(x) = \frac{1}{\Gamma(b)} a^b x^{b-1} e^{-ax} \ (0 \le x \le \infty)$ (A-8) where: $\mu = b / a$; and $\sigma^2 = b / a^2$ $\Gamma(x) = \int_0^t e^{-t} t^{x-1} dt \ for\ x > 0$ (A-9)
Exponential distribution	The exponential distribution is used as a probability model for the time between events occurring at random, or the time to failure of simple electronic devices. The probability density function of the exponential distribution is maximum when X = zero, then decays exponentially with increasing X. The probability density function of the exponential distribution can be expressed as in equation (A-10).	$f(x) = \lambda e^{-\lambda x} \ (o \le x \le \infty)$ (A-10) where: $\mu = 1/\lambda$; and $\sigma^2 = 1/\lambda^2$
Poisson distribution	The Poisson distribution is used as a probability model for the occurrence of rare events. The Poisson distribution is used in quality control, reliability, and queuing theory. When X indicates the number of events in a unit interval of time, X is called the Poisson random variable with mean number of events λ in a unit interval of time. The probability mass function of the Poisson distribution with parameter λ can be written as in equation (A-11).	$f(x) = \frac{e^{-\lambda} \lambda^x}{x!} \ (x = 0, 1, 2,...)$ (A-11) where: $\mu = \lambda$; and $\sigma^2 = \lambda$

Appendix B - NATO Supply Group

Table B-2 NATO Supply Group

Group Number	NATO Supply Group title
10	Weapons
11	Nuclear Ordinance
12	Fire Control Equipment
13	Ammunition and Explosives
14	Guided Missiles
15	Aircraft and Airframe Structural Components
16	Aircraft Components and Accessories
17	Aircraft Launching, Landing, and Ground Handling Equipment
18	Space Vehicles
19	Ships, Small Craft, Pontoons, and Floating Docks
20	Ship and Marine Equipment
22	Railway Equipment
23	Ground Effects Vehicles, Motor Vehicles, Trailers, and Cycles
24	Tractors
25	Vehicular Equipment Components
26	Tires and Tubes
28	Engines, Turbines and Components
29	Engine Accessories
30	Mechanical Power Transmission Equipment
31	Bearings
32	Woodworking Machinery and Equipment
34	Metalworking Machinery
35	Service and Trade Equipment
36	Special Industry Machinery
37	Agricultural Machinery and Equipment
38	Construction, Mining, Excavating, and Highway Maintenance Equipment
39	Materials Handling Equipment
40	Rope, Cable, Chain, and Fittings
41	Refrigeration, Air-Conditioning, and Air Circulating Equipment
42	Fire Fighting, Rescue and Safety Equipment
43	Pumps and Compressors
44	Furnace, Steam Plant, Drying Equipment; and Nuclear Reactors
45	Plumbing, Heating, and Sanitation Equipment
46	Water Purification and Sewage Treatment Equipment
47	Pipe, Tubing, Hose and Fittings
48	Valves
49	Maintenance and Repair Shop Equipment
51	Hand Tools

(*Continued*)

Table B-1 Continued

Group Number	NATO Supply Group title
52	Measuring Tools
53	Hardware and Abrasives
54	Prefabricated Structures and Scaffolding
55	Lumber, Millwork, Plywood and Veneer
56	Construction and Building Materials
58	Communication, Detection and Coherent Radiation Equipment
59	Electrical and Electronic Equipment Components
60	Fiber Optics Materials, Components, Assemblies, Accessories
61	Electric Wire and Power and Distribution Equipment
62	Lighting Fixtures and Lamps
63	Alarm, Signal and Security Detection Systems
65	Medical, Dental, and Veterinary Equipment and Supplies
66	Instruments and Laboratory Equipment
67	Photographic Equipment
68	Chemicals and Chemical Products
69	Training Aids and Devices
70	General Purpose Automatic Data Processing Equipment (Including Firmware), Software, Supplies and Support Equipment
71	Furniture
72	Household and Commercial Furnishings and Appliances
73	Food Preparation and Serving Equipment
74	Office Machines, Text Processing Systems, and Visible Record Equipment
75	Office Supplies and Devices
76	Books, Maps and Other Publications
77	Musical Instruments, Phonographs, and Home-Type Radios
78	Recreational and Athletic Equipment
79	Cleaning Equipment and Supplies
80	Brushes, Paints, Sealers and Adhesives
81	Containers, Packaging, and Packing Supplies
83	Textiles, Leathers, Furs, Apparel and Shoe Findings, Tents and Flags
84	Clothing, Individual Equipment, and Insignia
85	Toiletries
87	Agricultural Supplies
88	Live Animals
89	Subsistence
91	Fuels, Lubricants, Oils and Waxes
93	Nonmetallic Fabricated Materials
94	Nonmetallic Crude Materials
95	Metal Bars, Sheets, and Shapes
96	Ores, Minerals, and Their Primary Products
99	Miscellaneous

Appendix C - Time Series Plots Of The Sums Of The Monthly Aggregated Time Series Of The Spare Parts Demand For The Eight Pieces Of Equipment

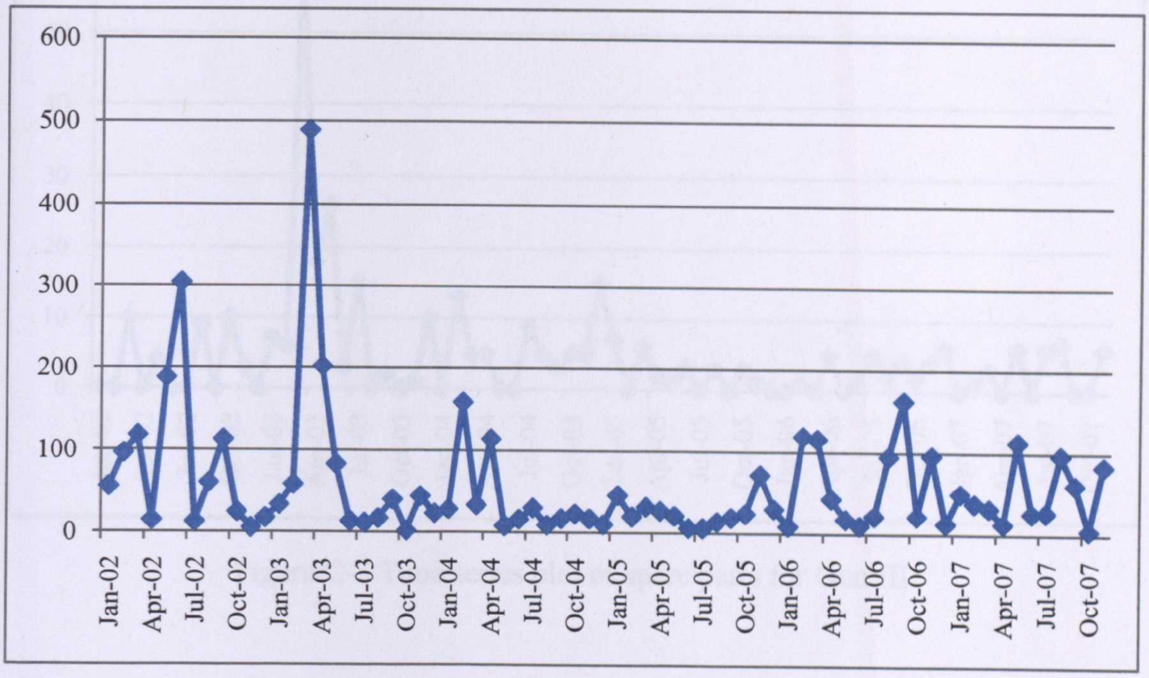


Figure C-1 Time series plot of spare parts for Gun I

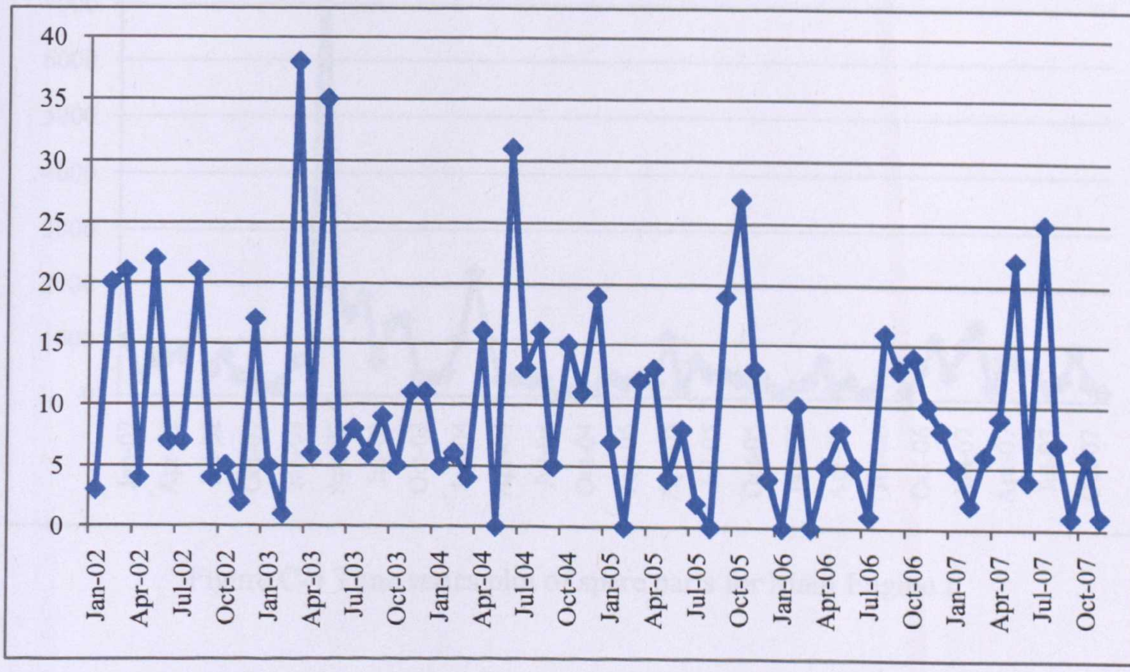


Figure C-2 Time series plot of spare parts for Gun II

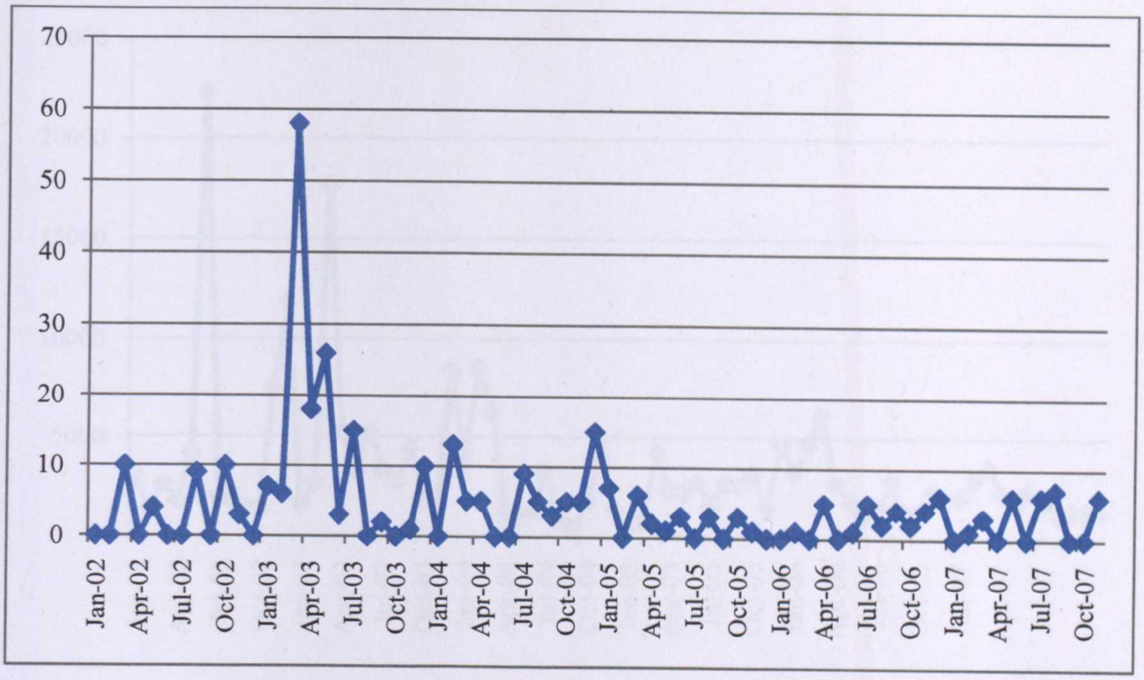


Figure C-3 Time series plot of spare parts for Gun III

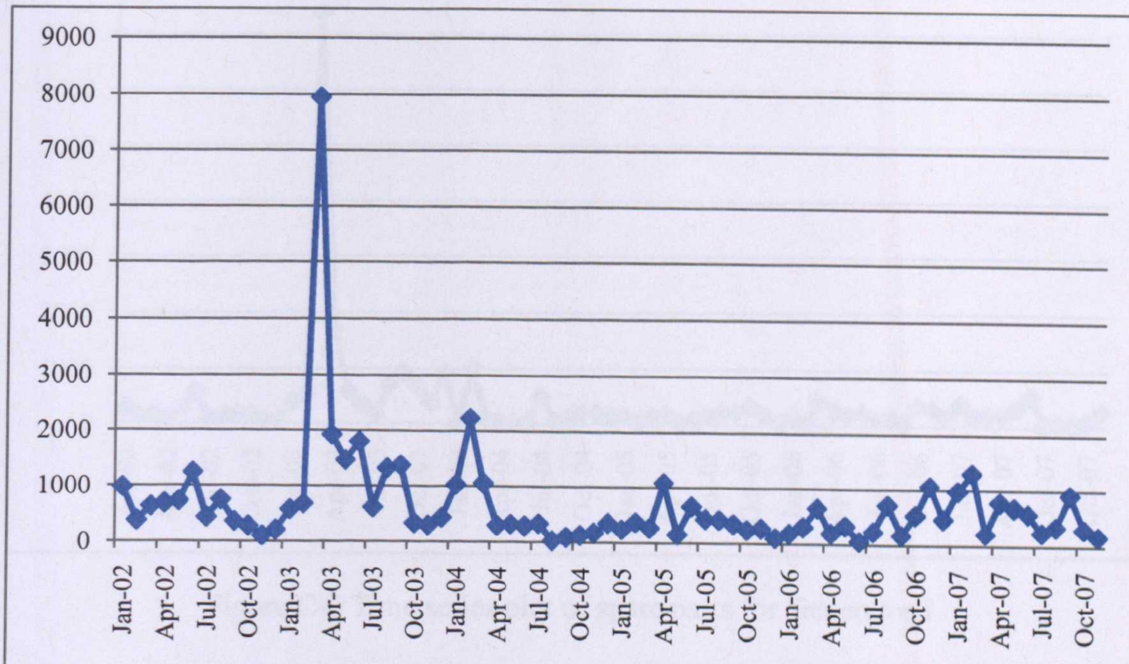


Figure C-4 Time series plot of spare parts for Main Engine I

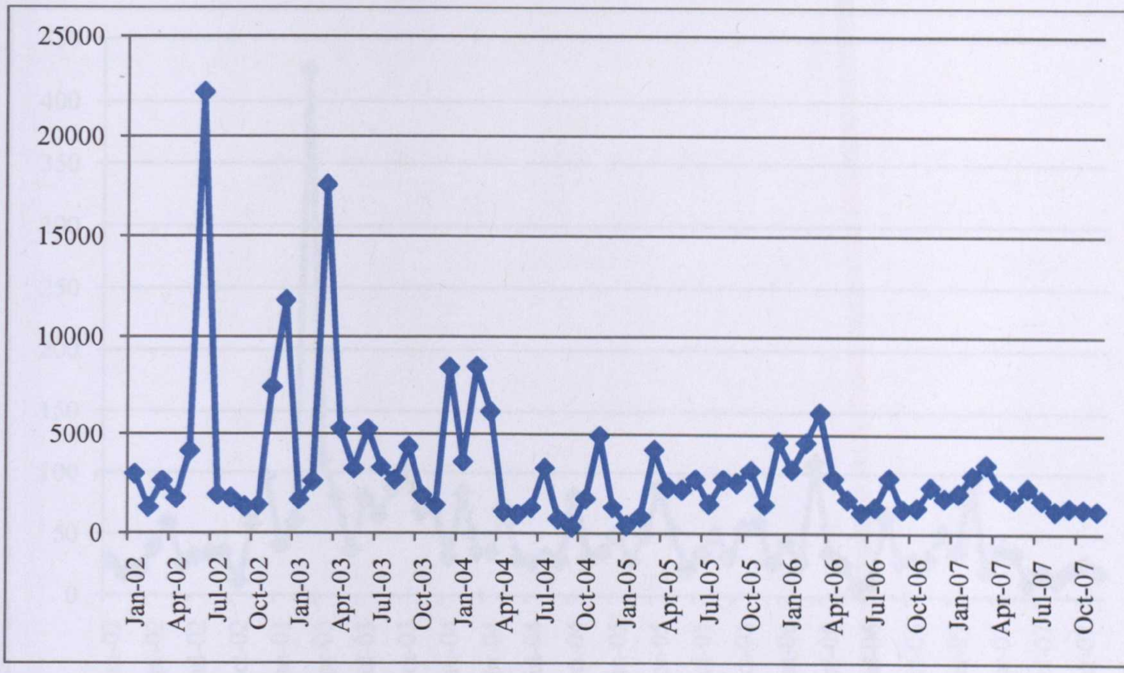


Figure C-5 Time series plot of spare parts for Main Engine II

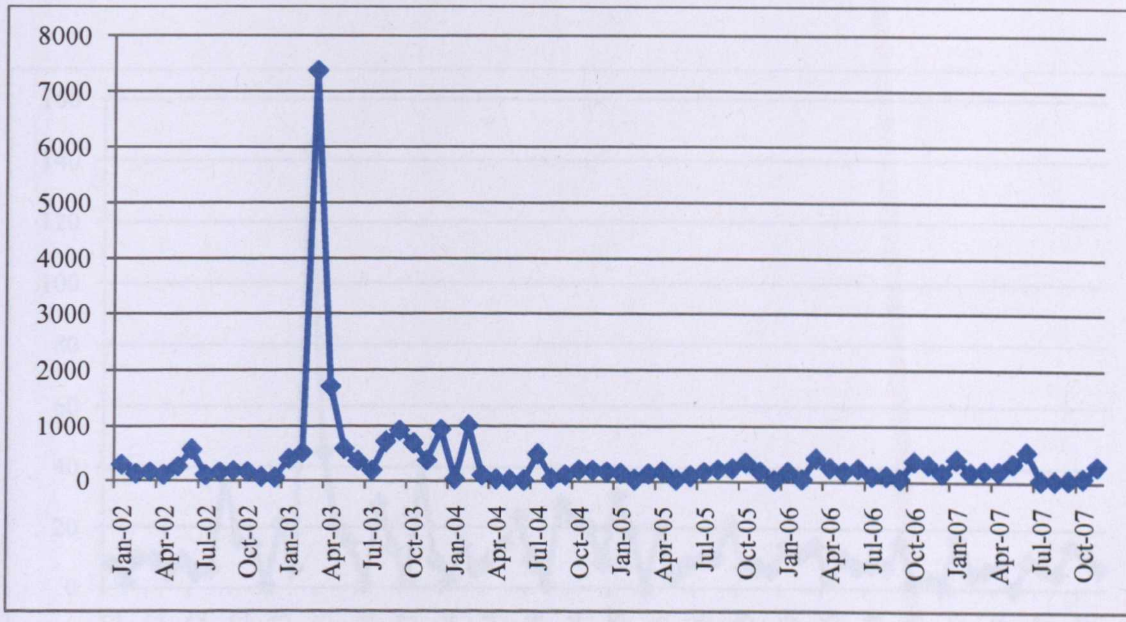


Figure C-6 Time series plot of spare parts for Generator I

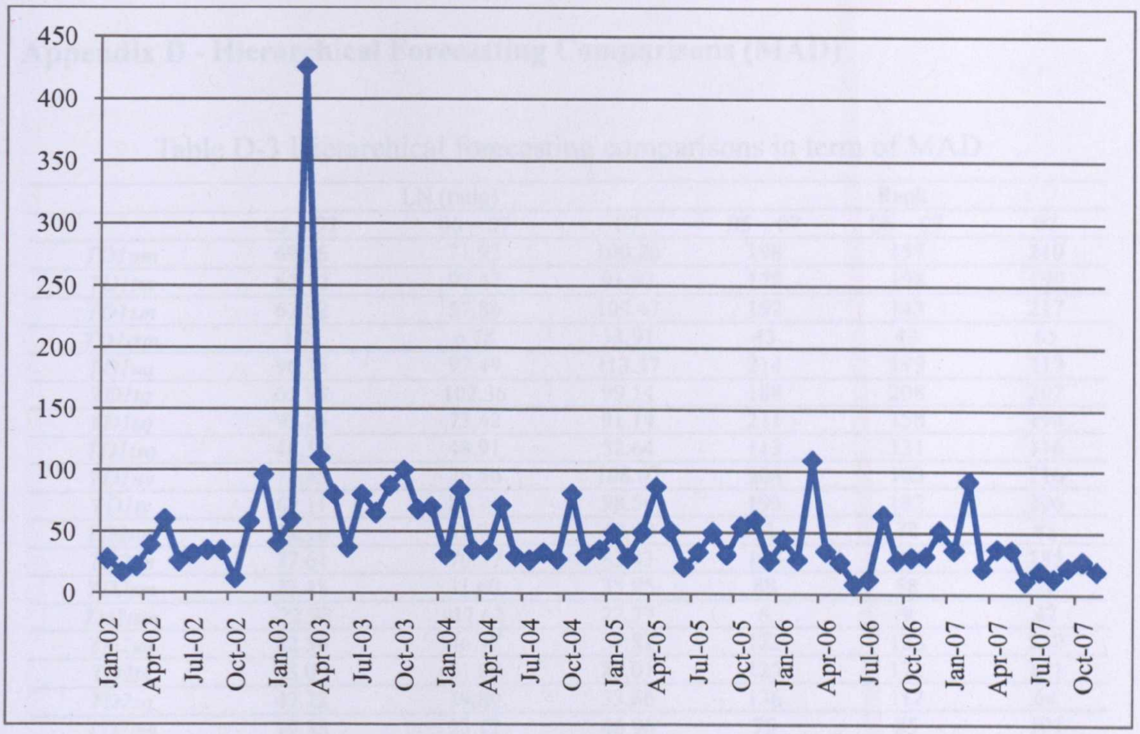


Figure C-7 Time series plot of spare parts for Air Compressor I

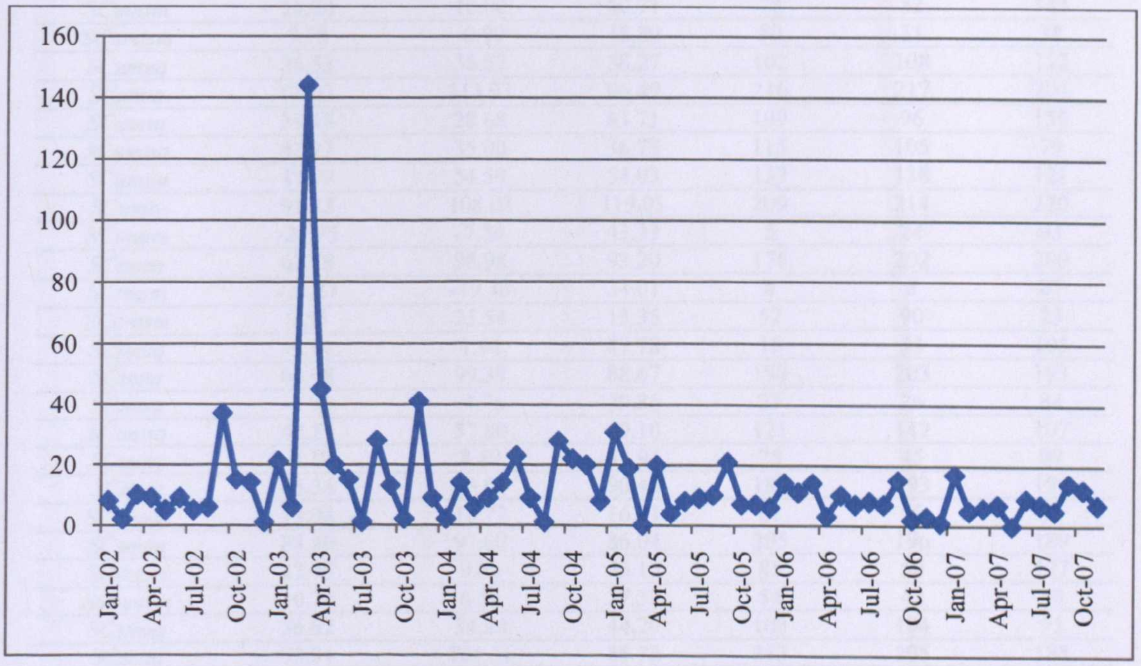


Figure C-8 Time series plot of spare parts for Radar I

Appendix D - Hierarchical Forecasting Comparisons (MAD)

Table D-3 Hierarchical forecasting comparisons in term of MAD

	LN (ratio)[2]			Rank		
	05 ~ 07	06 ~ 07	07	05 ~ 07	06 ~ 07	07
TD1um	69.36	71.92	100.26	198	157	210
TD1tm	63.42	97.47	91.93	178	198	199
TD1sm	67.02	57.86	109.47	192	143	217
TD1tsm	3.92	6.78	33.91	43	43	65
TD1uq	96.33	97.49	113.37	214	199	219
TD1tq	65.93	102.36	99.14	188	208	207
TD1sq	93.25	73.42	91.14	211	158	198
TD1tsq	40.59	48.91	52.64	113	131	116
TD1uy	78.45	76.56	108.07	204	163	216
TD1ty	67.11	86.74	98.59	193	182	206
TD2um	18.16	19.73	40.40	71	78	88
TD2tm	37.61	70.57	83.43	104	155	185
TD2sm	23.48	11.60	35.95	88	58	76
TD2tsm	-22.98	-17.62	22.77	6	9	42
TD2uq	42.56	46.97	57.82	122	128	130
TD2tq	43.09	81.00	88.01	125	172	191
TD2sq	47.72	39.80	33.86	136	112	64
TD2tsq	19.38	23.12	46.96	75	85	101
TD2uy	29.29	34.30	59.98	92	103	139
TD2ty	42.85	68.25	99.91	123	151	209
SCumum	18.35	19.50	34.53	72	77	69
SCumtm	86.89	107.57	100.63	207	213	212
SCumsm	20.00	10.00	60.71	78	52	143
SCumtsm	7.78	-0.99	18.89	50	31	38
SCumuq	36.53	36.52	58.27	102	108	132
SCumtq	98.02	113.93	96.49	216	217	203
SCumsq	39.48	28.68	63.71	109	96	150
SCumtsq	41.27	35.00	36.79	115	105	79
SCumuy	47.99	54.59	54.03	137	138	121
SCumty	91.32	108.02	119.03	209	214	220
SCtmum	-24.75	-7.50	43.33	5	24	93
SCtmtm	62.78	98.98	93.20	176	202	200
SCtmsm	-21.87	-19.40	34.01	8	8	67
SCtmtsm	9.55	25.54	13.35	52	90	23
SCtmuq	-9.14	1.94	47.78	16	33	105
SCtmtq	66.08	99.39	88.67	190	203	193
SCtmsq	-2.96	-4.76	39.86	27	26	84
SCtmtsq	42.25	57.80	49.10	121	142	107
SCtmuy	-4.19	8.49	45.91	25	45	99
SCtmty	65.36	96.64	90.42	187	193	195
SCsmum	14.08	10.72	16.81	61	56	32
SCsmtm	84.80	97.00	86.04	205	196	189
SCsmsm	22.95	9.65	56.13	85	50	127
SCsmtsm	10.74	6.54	37.32	54	42	80
SCsmuq	36.62	34.80	44.20	103	104	95
SCsmtq	95.84	101.31	85.72	213	205	188

(*Continued*)

[2] LN(ratio) = the sum of natural log relative error [ln($error_{HF}$/$error_{DF}$)] for MAD over the 300 items.

Table D-1 Continued

	LN (ratio)			Rank		
	05 ~ 07	06 ~ 07	07	05 ~ 07	06 ~ 07	07
SCsmsq	40.44	31.82	47.28	112	98	103
SCsmtsq	39.37	29.02	36.78	108	97	78
SCsmuy	48.59	53.10	50.82	140	135	110
SCsmty	91.37	105.54	99.58	210	212	208
SCtsmum	-26.02	-27.74	2.79	3	1	8
SCtsmtm	58.63	78.87	74.55	167	170	172
SCtsmsm	-20.63	-24.98	15.64	9	3	28
SCtsmtsm	-3.05	5.16	2.77	26	38	7
SCtsmuq	-5.92	-1.82	13.09	21	29	22
SCtsmtq	65.18	84.77	74.79	186	179	174
SCtsmsq	-0.70	-8.35	11.20	32	19	18
SCtsmtsq	25.17	25.14	26.73	90	89	51
SCtsmuy	3.23	10.47	35.54	40	54	73
SCtsmty	56.92	81.23	78.85	164	173	179
SCuqum	29.60	32.78	53.39	93	99	120
SCuqtm	96.43	112.26	100.53	215	215	211
SCuqsm	33.62	32.79	69.24	97	100	163
SCuqtsm	20.05	23.47	34.93	79	87	71
SCuquq	48.09	53.38	66.17	138	137	159
SCuqtq	106.02	117.62	103.50	220	219	214
SCuqsq	52.33	48.46	62.55	153	129	148
SCuqtsq	53.52	52.20	45.14	156	133	98
SCuquy	60.90	68.87	64.75	173	152	154
SCuqty	99.52	118.07	102.63	217	220	213
SCtqum	-32.10	-12.71	33.56	1	13	62
SCtqtm	59.20	96.91	86.77	170	195	190
SCtqsm	-25.25	-23.09	24.69	4	6	47
SCtqtsm	9.64	28.45	15.93	53	95	29
SCtquq	-15.07	-4.77	34.85	13	25	70
SCtqtq	63.52	100.46	91.13	179	204	197
SCtqsq	-6.83	-8.95	37.86	20	18	81
SCtqtsq	41.88	59.29	45.02	118	144	97
SCtquy	-8.79	5.20	35.54	17	40	74
SCtqty	62.85	98.63	88.95	177	201	194
SCsqum	24.38	22.66	32.59	89	83	61
SCsqtm	94.86	105.50	88.62	212	211	192
SCsqsm	33.37	25.57	58.94	96	91	136
SCsqtsm	22.60	15.90	31.61	84	71	57
SCsquq	43.44	40.57	43.79	126	116	94
SCsqtq	104.83	113.68	81.76	219	216	183
SCsqsq	48.70	41.78	51.64	143	121	113
SCsqtsq	51.79	41.76	36.35	149	120	77
SCsquy	57.11	65.42	52.65	165	148	117
SCsqty	100.70	115.70	98.01	218	218	205
SCtsqum	-27.00	-20.55	7.75	2	7	10
SCtsqtm	55.88	74.32	70.60	161	160	166
SCtsqsm	-18.54	-23.69	15.28	10	5	25
SCtsqtsm	6.75	10.59	9.30	48	55	13
SCtsquq	-5.10	-1.34	13.52	23	30	24
SCtsqtq	61.22	82.75	68.22	174	177	161
SCtsqsq	-0.39	-7.95	22.62	33	21	41
SCtsqtsq	35.76	33.99	44.87	99	102	96
SCtsquy	3.37	15.98	35.43	41	72	72
SCtsqty	58.92	81.84	77.76	169	175	177
SCuyum	17.68	14.68	38.44	69	65	82
SCuytm	78.28	97.12	95.94	203	197	202
SCuysm	19.41	17.13	54.06	76	74	122
SCuytsm	13.41	9.39	35.81	57	49	75

(*Continued*)

Table D-1 Continued

	LN (ratio)			Rank		
	05 ~ 07	06 ~ 07	07	05 ~ 07	06 ~ 07	07
SCuyuq	35.77	38.27	41.77	100	110	91
SCuytq	87.83	102.68	109.68	208	209	218
SCuysq	42.14	37.29	56.75	120	109	128
SCuytsq	38.20	39.75	55.44	106	111	125
SCuyuy	47.11	56.08	64.20	133	141	153
SCuyty	84.91	103.69	106.19	206	210	215
SCtyum	-12.35	-23.79	26.86	14	4	52
SCtytm	64.96	87.68	79.99	185	183	181
SCtysm	-5.39	-25.56	-0.91	22	2	5
SCtytsm	15.98	20.55	18.02	63	82	37
SCtyuq	3.76	-7.58	23.41	42	23	44
SCtytq	69.68	91.62	75.88	199	188	176
SCtysq	13.65	-11.74	23.45	58	15	45
SCtytsq	44.00	49.01	42.52	129	132	92
SCtyuy	13.23	4.14	30.80	56	35	55
SCtyty	66.36	89.55	65.48	191	187	156
WCumum	16.62	20.43	41.03	66	81	89
WCumtm	48.68	76.54	74.66	142	162	173
WCumsm	23.17	15.74	62.54	86	69	147
WCumtsm	-0.08	5.37	8.04	34	41	11
WCumuq	36.22	40.52	58.56	101	115	134
WCumtq	53.74	78.64	85.45	158	168	187
WCumsq	40.35	33.42	63.59	111	101	149
WCumtsq	41.30	43.18	33.95	116	122	66
WCumuy	41.59	39.84	40.38	117	113	87
WCumty	64.18	77.95	59.22	181	166	137
WCtmum	-10.06	-3.32	11.38	15	27	19
WCtmtm	64.53	101.67	94.23	183	206	201
WCtmsm	0.49	-8.15	15.39	36	20	26
WCtmtsm	5.46	16.54	15.94	45	73	30
WCtmuq	6.89	15.80	32.33	49	70	59
WCtmtq	67.34	102.18	97.93	195	207	204
WCtmsq	21.17	11.17	17.32	82	57	35
WCtmtsq	56.59	66.33	62.40	163	149	146
WCtmuy	23.38	12.74	47.08	87	62	102
WCtmty	74.83	97.92	73.78	202	200	169
WCsmum	13.73	14.04	34.28	59	64	68
WCsmtm	43.50	67.53	61.43	127	150	145
WCsmsm	26.08	15.73	60.37	91	68	140
WCsmtsm	0.25	5.17	17.04	35	39	34
WCsmuq	34.80	36.50	57.29	98	107	129
WCsmtq	52.52	70.45	65.49	154	154	157
WCsmsq	42.07	35.72	58.34	119	106	133
WCsmtsq	43.69	43.25	40.21	128	123	85
WCsmuy	40.30	40.27	41.71	110	114	90
WCsmty	64.74	76.78	53.18	184	164	119
WCtsmum	-15.46	-12.02	-1.57	12	14	4
WCtsmtm	48.22	77.16	75.10	139	165	175
WCtsmsm	-0.78	-9.28	26.51	31	17	49
WCtsmtsm	-2.43	4.30	11.46	28	36	20
WCtsmuq	4.62	13.59	23.30	44	63	43
WCtsmtq	52.96	81.36	73.06	155	174	168
WCtsmsq	17.77	9.19	17.04	70	47	33
WCtsmtsq	45.20	46.60	47.40	130	127	104
WCtsmuy	19.21	12.73	31.37	74	61	56
WCtsmty	63.65	78.55	51.71	180	167	114
WCuqum	22.40	27.77	50.36	83	93	108

(Continued)

Table D-1 Continued

	LN (ratio)			Rank		
	05 ~ 07	06 ~ 07	07	05 ~ 07	06 ~ 07	07
WCuqtm	51.28	78.71	73.96	147	169	171
WCuqsm	30.65	26.24	64.17	94	92	152
WCuqtsm	6.55	15.61	9.70	47	67	14
WCuquq	46.09	53.16	72.78	131	136	167
WCuqtq	58.86	82.11	85.38	168	176	186
WCuqsq	49.29	44.50	59.89	144	126	138
WCuqtsq	46.61	48.72	28.19	132	130	53
WCuquy	51.90	55.46	57.92	151	140	131
WCuqty	68.64	87.88	61.19	197	184	144
WCtqum	-22.34	-15.87	-0.16	7	10	6
WCtqtm	56.44	95.13	80.37	162	192	182
WCtqsm	-7.55	-15.22	-7.28	19	11	2
WCtqtsm	1.80	15.03	10.93	37	66	17
WCtquq	-1.09	4.62	17.89	30	37	36
WCtqtq	61.37	96.76	91.08	175	194	196
WCtqsq	13.89	3.90	10.11	60	34	15
WCtqtsq	52.15	60.15	52.68	152	145	118
WCtquy	16.21	9.19	33.75	64	48	63
WCtqty	70.73	94.24	67.47	200	191	160
WCsqum	20.89	20.04	40.35	80	79	86
WCsqtm	51.66	69.43	51.76	148	153	115
WCsqsm	32.42	23.18	58.87	95	86	135
WCsqtsm	9.47	12.34	19.45	51	60	39
WCsquq	42.86	43.58	60.48	124	124	141
WCsqtq	59.93	73.80	68.67	172	159	162
WCsqsq	48.66	41.19	60.60	141	118	142
WCsqtsq	47.49	44.12	25.67	135	125	48
WCsquy	47.33	52.85	51.50	134	134	112
WCsqty	68.28	84.87	55.16	196	180	124
WCtsqum	-17.85	-14.54	-6.29	11	12	3
WCtsqtm	54.25	85.69	73.88	159	181	170
WCtsqsm	-5.10	-9.50	10.64	24	16	16
WCtsqtsm	3.00	9.96	6.19	39	51	9
WCtsquq	5.61	12.02	19.54	46	59	40
WCtsqtq	59.65	89.48	81.76	171	186	184
WCtsqsq	16.77	10.44	15.94	67	53	31
WCtsqtsq	51.88	54.71	47.82	150	139	106
WCtsquy	18.50	17.98	32.54	73	75	60
WCtsqty	66.06	84.73	63.97	189	178	151
WCuyum	16.32	9.08	31.94	65	46	58
WCuytm	50.01	70.84	69.39	145	156	164
WCuysm	19.56	8.03	45.93	77	44	100
WCuytsm	2.53	-0.92	24.15	38	32	46
WCuyuq	37.84	41.38	54.36	105	119	123
WCuytq	58.46	74.59	79.18	166	161	180
WCuysq	38.61	27.94	51.39	107	94	111
WCuytsq	41.03	40.69	38.83	114	117	83
WCuyuy	50.32	60.36	65.77	146	146	158
WCuyty	54.81	80.20	56.06	160	171	126
WCtyum	-7.61	-3.12	-14.63	18	28	1
WCtytm	64.48	88.68	70.31	182	185	165
WCtysm	-2.07	-7.78	15.59	29	22	27
WCtytsm	16.81	22.73	8.53	68	84	12
WCtyuq	13.06	20.19	11.61	55	80	21
WCtytq	73.96	91.97	78.31	201	189	178
WCtysq	20.91	18.57	28.49	81	76	54
WCtytsq	53.65	62.37	50.79	157	147	109
WCtyuy	14.55	24.35	26.66	62	88	50
WCtyty	67.30	92.08	65.17	194	190	155

Appendix E - Hierarchical Forecasting Comparisons (RMSE)

Table E-4 Hierarchical forecasting comparisons in term of RMSE

	LN (ratio)[3]			Rank		
	05 ~ 07	06 ~ 07	07	05 ~ 07	06 ~ 07	07
TD1um	61.86	68.97	100.26	185	152	210
TD1tm	56.70	96.11	91.93	172	192	199
TD1sm	62.48	54.66	109.47	188	138	217
TD1tsm	2.49	6.87	33.91	43	43	65
TD1uq	89.02	96.80	113.37	209	194	219
TD1tq	58.21	100.22	99.14	176	202	207
TD1sq	89.49	73.54	91.14	210	157	198
TD1tsq	38.72	49.65	52.64	118	132	116
TD1uy	72.04	73.96	108.07	201	159	216
TD1ty	64.86	85.34	98.59	190	179	206
TD2um	12.14	18.38	40.40	60	78	88
TD2tm	32.15	68.58	83.43	102	151	185
TD2sm	18.29	10.15	35.95	81	53	76
TD2tsm	-23.81	-19.41	22.77	8	9	42
TD2uq	36.70	46.71	57.82	110	128	130
TD2tq	34.63	78.45	88.01	105	166	191
TD2sq	43.82	41.62	33.86	137	116	64
TD2tsq	16.69	22.62	46.96	74	84	101
TD2uy	22.11	30.63	59.98	91	97	139
TD2ty	39.98	66.06	99.91	126	149	209
SCumum	13.51	17.79	34.53	64	76	69
SCumtm	83.75	108.39	100.63	207	213	212
SCumsm	16.20	7.31	60.71	72	46	143
SCumtsm	7.10	0.21	18.89	51	33	38
SCumuq	31.51	35.55	58.27	101	106	132
SCumtq	94.41	114.88	96.49	215	216	203
SCumsq	36.20	27.21	63.71	108	94	150
SCumtsq	42.33	39.38	36.79	133	110	79
SCumuy	42.24	54.97	54.03	131	139	121
SCumty	90.18	109.10	119.03	212	214	220
SCtmum	-28.76	-11.85	43.33	3	16	93
SCtmtm	58.79	99.38	93.20	178	200	200
SCtmsm	-25.35	-23.93	34.01	6	7	67
SCtmtsm	8.34	26.44	13.35	53	93	23
SCtmuq	-12.37	-0.22	47.78	14	32	105
SCtmtq	61.14	100.05	88.67	183	201	193
SCtmsq	-5.80	-8.14	39.86	25	23	84
SCtmtsq	39.83	58.99	49.10	124	144	107

(*Continued*)

[3] LN(ratio) = the sum of natural log relative error [$\ln(error_{HF}/error_{DF})$] for RMSE over the 300 items.

Table E-1 Continued

	LN (ratio)			Rank		
	05 ~ 07	06 ~ 07	07	05 ~ 07	06 ~ 07	07
SCtmuy	-7.50	7.26	45.91	22	45	99
SCtmty	63.50	97.17	90.42	189	196	195
SCsmum	10.48	10.33	16.81	56	55	32
SCsmtm	83.75	100.48	86.04	206	203	189
SCsmsm	19.10	8.05	56.13	83	49	127
SCsmtsm	8.74	7.56	37.32	54	47	80
SCsmuq	33.08	34.64	44.20	103	103	95
SCsmtq	94.50	104.89	85.72	216	210	188
SCsmsq	37.27	32.41	47.28	113	101	103
SCsmtsq	40.69	31.62	36.78	127	100	78
SCsmuy	42.32	52.88	50.82	132	135	110
SCsmty	89.70	106.40	99.58	211	211	208
SCtsmum	-27.68	-27.80	2.79	5	1	8
SCtsmtm	56.04	81.61	74.55	170	172	172
SCtsmsm	-22.71	-24.93	15.64	9	5	28
SCtsmtsm	-4.26	5.25	2.77	27	38	7
SCtsmuq	-8.04	-0.40	13.09	20	31	22
SCtsmtq	61.10	86.56	74.79	182	180	174
SCtsmsq	-1.26	-6.92	11.20	34	25	18
SCtsmtsq	24.51	25.56	26.73	92	89	51
SCtsmuy	-0.26	11.80	35.54	38	59	73
SCtsmty	58.22	83.94	78.85	177	176	179
SCuqum	25.38	31.48	53.39	93	99	120
SCuqtm	92.65	113.78	100.53	214	215	211
SCuqsm	27.96	31.12	69.24	95	98	163
SCuqtsm	17.77	25.09	34.93	80	87	71
SCuquq	42.21	52.37	66.17	130	134	159
SCuqtq	103.20	119.12	103.50	220	219	214
SCuqsq	48.22	48.32	62.55	148	129	148
SCuqtsq	52.48	56.30	45.14	160	141	98
SCuquy	54.51	70.36	64.75	166	153	154
SCuqty	98.55	119.83	102.63	217	220	213
SCtqum	-36.38	-16.20	33.56	1	11	62
SCtqtm	53.46	97.95	86.77	164	197	190
SCtqsm	-28.81	-26.74	24.69	2	3	47
SCtqtsm	6.44	28.93	15.93	50	96	29
SCtquq	-17.02	-5.16	34.85	13	26	70
SCtqtq	57.97	100.54	91.13	175	204	197
SCtqsq	-9.25	-12.30	37.86	18	15	81
SCtqtsq	38.37	59.80	45.02	117	146	97
SCtquy	-12.13	5.61	35.54	17	41	74
SCtqty	60.64	98.79	88.95	180	199	194
SCsqum	21.37	22.89	32.59	89	85	61
SCsqtm	91.46	107.66	88.62	213	212	192
SCsqsm	30.03	25.93	58.94	97	92	136
SCsqtsm	20.90	16.28	31.61	87	74	57
SCsquq	39.97	42.34	43.79	125	119	94
SCsqtq	102.14	116.01	81.76	219	217	183

(*Continued*)

Table E-1 Continued

	LN (ratio)			Rank		
	05 ~ 07	06 ~ 07	07	05 ~ 07	06 ~ 07	07
SCsqsq	45.87	43.58	51.64	144	123	113
SCsqtsq	52.06	45.54	36.35	158	126	77
SCsquy	52.16	68.06	52.65	159	150	117
SCsqty	98.74	116.76	98.01	218	218	205
SCtsqum	-28.66	-21.76	7.75	4	8	10
SCtsqtm	52.65	75.01	70.60	161	161	166
SCtsqsm	-20.30	-24.69	15.28	11	6	25
SCtsqtsm	6.22	11.80	9.30	49	60	13
SCtsquq	-6.64	-0.75	13.52	24	30	24
SCtsqtq	57.46	83.37	68.22	173	175	161
SCtsqsq	-0.90	-8.70	22.62	36	22	41
SCtsqtsq	35.60	35.49	44.87	107	105	96
SCtsquy	-1.23	15.59	35.43	35	70	72
SCtsqty	59.11	84.61	77.76	179	177	177
SCuyum	13.40	13.06	38.44	63	65	82
SCuytm	76.49	97.98	95.94	203	198	202
SCuysm	16.17	14.41	54.06	71	67	122
SCuytsm	11.75	8.58	35.81	59	50	75
SCuyuq	30.65	38.26	41.77	100	109	91
SCuytq	84.69	103.07	109.68	208	208	218
SCuysq	37.71	35.55	56.75	115	107	128
SCuytsq	37.33	42.28	55.44	114	118	125
SCuyuy	39.19	54.50	64.20	121	137	153
SCuyty	82.45	103.42	106.19	205	209	215
SCtyum	-12.33	-26.28	26.86	15	4	52
SCtytm	61.17	87.25	79.99	184	181	181
SCtysm	-3.34	-27.21	-0.91	29	2	5
SCtytsm	16.82	20.96	18.02	77	80	37
SCtyuq	4.78	-8.75	23.41	47	21	44
SCtytq	66.02	92.74	75.88	191	189	176
SCtysq	15.48	-12.40	23.45	69	14	45
SCtytsq	45.41	49.46	42.52	142	131	92
SCtyuy	12.79	4.11	30.80	62	34	55
SCtyty	67.14	91.90	65.48	195	188	156
WCumum	11.15	17.81	41.03	57	77	89
WCumtm	46.48	76.68	74.66	145	163	173
WCumsm	18.60	12.16	62.54	82	61	147
WCumtsm	-1.93	5.41	8.04	32	39	11
WCumuq	30.63	40.64	58.56	99	114	134
WCumtq	49.53	78.13	85.45	152	165	187
WCumsq	36.96	33.04	63.59	111	102	149
WCumtsq	39.09	42.47	33.95	119	120	66
WCumuy	37.12	41.05	40.38	112	115	87
WCumty	66.41	80.07	59.22	192	168	137
WCtmum	-12.32	-4.10	11.38	16	27	19
WCtmtm	60.69	101.18	94.23	181	206	201
WCtmsm	-1.80	-9.60	15.39	33	20	26
WCtmtsm	4.16	17.20	15.94	45	75	30

(Continued)

Table E-1 Continued

	LN (ratio)			Rank		
	05 ~ 07	06 ~ 07	07	05 ~ 07	06 ~ 07	07
WCtmuq	4.25	16.02	32.33	46	72	59
WCtmtq	62.36	101.89	97.93	187	207	204
WCtmsq	20.50	12.31	17.32	86	62	35
WCtmtsq	53.37	65.65	62.40	163	148	146
WCtmuy	21.67	10.10	47.08	90	51	102
WCtmty	76.79	100.71	73.78	204	205	169
WCsmum	9.77	12.63	34.28	55	63	68
WCsmtm	42.49	70.46	61.43	134	154	145
WCsmsm	21.17	12.90	60.37	88	64	140
WCsmtsm	-0.37	5.05	17.04	37	37	34
WCsmuq	30.40	36.64	57.29	98	108	129
WCsmtq	50.43	73.58	65.49	154	158	157
WCsmsq	39.15	35.34	58.34	120	104	133
WCsmtsq	42.14	43.48	40.21	129	122	85
WCsmuy	35.08	40.18	41.71	106	112	90
WCsmty	66.84	78.52	53.18	193	167	119
WCtsmum	-17.08	-11.78	-1.57	12	17	4
WCtsmtm	44.44	77.79	75.10	138	164	175
WCtsmsm	-4.74	-11.18	26.51	26	18	49
WCtsmtsm	-3.78	4.23	11.46	28	35	20
WCtsmuq	1.10	14.89	23.30	40	68	43
WCtsmtq	48.12	80.91	73.06	147	170	168
WCtsmsq	16.81	10.25	17.04	76	54	33
WCtsmtsq	43.26	46.21	47.40	136	127	104
WCtsmuy	15.84	11.79	31.37	70	58	56
WCtsmty	67.05	82.94	51.71	194	174	114
WCuqum	17.10	25.78	50.36	78	91	108
WCuqtm	48.65	80.81	73.96	149	169	171
WCuqsm	25.82	24.37	64.17	94	86	152
WCuqtsm	4.98	15.82	9.70	48	71	14
WCuquq	39.32	52.07	72.78	122	133	167
WCuqtq	54.97	82.77	85.38	167	173	186
WCuqsq	44.96	44.33	59.89	139	125	138
WCuqtsq	45.09	48.44	28.19	141	130	53
WCuquy	45.02	56.53	57.92	140	142	131
WCuqty	69.62	90.89	61.19	199	186	144
WCtqum	-24.34	-16.53	-0.16	7	10	6
WCtqtm	51.74	94.96	80.37	157	190	182
WCtqsm	-8.99	-15.29	-7.28	19	12	2
WCtqtsm	0.40	15.06	10.93	39	69	17
WCtquq	-2.79	5.51	17.89	30	40	36
WCtqtq	55.85	96.18	91.08	169	193	196
WCtqsq	13.52	4.35	10.11	65	36	15
WCtqtsq	49.88	59.42	52.68	153	145	118
WCtquy	14.77	7.61	33.75	67	48	63
WCtqty	72.69	96.91	67.47	202	195	160
WCsqum	16.80	18.97	40.35	75	79	86

(Continued)

Table E-1 Continued

	LN (ratio)			Rank		
	05 ~ 07	06 ~ 07	07	05 ~ 07	06 ~ 07	07
WCsqtm	50.44	72.65	51.76	155	156	115
WCsqsm	27.99	21.25	58.87	96	81	135
WCsqtsm	7.62	11.02	19.45	52	56	39
WCsquq	38.19	43.45	60.48	116	121	141
WCsqtq	57.62	75.59	68.67	174	162	162
WCsqsq	45.84	41.86	60.60	143	117	142
WCsqtsq	46.94	44.10	25.67	146	124	48
WCsquy	41.13	53.59	51.50	128	136	112
WCsqty	69.57	88.54	55.16	198	184	124
WCtsqum	-20.80	-14.97	-6.29	10	13	3
WCtsqtm	48.93	85.23	73.88	151	178	170
WCtsqsm	-6.93	-9.85	10.64	23	19	16
WCtsqtsm	2.06	10.12	6.19	41	52	9
WCtsquq	2.08	13.37	19.54	42	66	40
WCtsqtq	54.12	88.17	81.76	165	182	184
WCtsqsq	16.29	11.26	15.94	73	57	31
WCtsqtsq	50.89	55.14	47.82	156	140	106
WCtsquy	14.83	16.17	32.54	68	73	60
WCtsqty	68.95	88.53	63.97	197	183	151
WCuyum	13.94	7.19	31.94	66	44	58
WCuytm	48.72	71.14	69.39	150	155	164
WCuysm	17.11	6.74	45.93	79	42	100
WCuytsm	4.11	-1.31	24.15	44	28	46
WCuyuq	33.41	39.72	54.36	104	111	123
WCuytq	56.45	74.95	79.18	171	160	180
WCuysq	36.29	28.58	51.39	109	95	111
WCuytsq	39.64	40.48	38.83	123	113	83
WCuyuy	42.81	58.77	65.77	135	143	158
WCuyty	55.64	81.15	56.06	168	171	126
WCtyum	-7.83	-0.87	-14.63	21	29	1
WCtytm	62.21	89.35	70.31	186	185	165
WCtysm	-2.53	-7.34	15.59	31	24	27
WCtytsm	19.36	25.60	8.53	84	90	12
WCtyuq	11.50	22.46	11.61	58	83	21
WCtytq	69.83	91.25	78.31	200	187	178
WCtysq	20.35	21.72	28.49	85	82	54
WCtytsq	52.99	63.19	50.79	162	147	109
WCtyuy	12.19	25.28	26.66	61	88	50
WCtyty	68.89	95.06	65.17	196	191	155

Appendix F - Hierarchical Forecasting Comparisons (2005 ~ 2007)

Table F-5 Hierarchical forecasting comparisons in term of the mean rank of MAD and RMSE

	MAD		RMSE	
	Mean rank[4]	Rank[5]	Mean rank	Rank
TD1um	115.80	122	116.08	124
TD1tm	121.82	153	120.45	145
TD1sm	115.49	120	115.57	121
TD1tsm	91.06	37	92.17	41
TD1uq	129.67	197	129.06	192
TD1tq	127.29	184	125.43	172
TD1sq	127.54	187	128.66	190
TD1tsq	110.91	103	111.33	106
TD1uy	123.18	165	120.63	147
TD1ty	126.89	182	126.70	180
TD2um	103.90	79	105.10	82
TD2tm	118.60	137	116.04	123
TD2sm	106.06	88	106.00	86
TD2tsm	84.90	17	85.26	19
TD2uq	118.15	134	119.38	140
TD2tq	122.44	157	120.02	144
TD2sq	121.16	148	122.07	151
TD2tsq	106.54	89	107.18	92
TD2uy	107.56	91	106.75	90
TD2ty	122.07	155	123.27	163
SCumum	103.23	76	104.37	80
SCumtm	126.30	179	126.86	182
SCumsm	103.38	77	104.32	79
SCumtsm	94.06	50	93.83	49
SCumuq	110.56	102	112.15	110
SCumtq	131.78	203	130.02	196
SCumsq	112.12	113	112.77	116
SCumtsq	110.35	100	110.98	105
SCumuy	108.52	95	107.62	95
SCumty	125.39	174	127.15	184
SCtmum	79.34	4	79.21	7
SCtmtm	132.75	205	131.76	204
SCtmsm	79.05	3	78.38	3
SCtmtsm	99.25	60	98.78	60
SCtmuq	84.92	18	83.93	17
SCtmtq	134.88	215	133.93	213

(Continued)

[4] Mean ranks for the forecasting method in terms of MAD or RMSE over the 300 items;

[5] The ranks of the forecasting method among the 220 hierarchical forecasting methods based on mean rank.

Table F-1 Continued

	MAD		RMSE	
	Mean rank	Rank	Mean rank	Rank
SCtmsq	89.02	26	88.25	26
SCtmtsq	117.59	131	116.14	125
SCtmuy	80.01	7	79.00	4
SCtmty	123.00	163	122.80	159
SCsmum	103.75	78	104.18	78
SCsmtm	125.83	177	127.04	183
SCsmsm	105.99	86	105.15	83
SCsmtsm	96.24	58	94.84	51
SCsmuq	111.58	108	112.32	111
SCsmtq	131.51	201	131.50	201
SCsmsq	116.25	124	116.02	122
SCsmtsq	111.01	104	112.69	115
SCsmuy	110.06	99	108.08	97
SCsmty	128.55	192	129.57	195
SCtsmum	78.70	2	79.07	5
SCtsmtm	122.57	159	122.53	155
SCtsmsm	81.30	10	81.03	9
SCtsmtsm	90.71	35	89.68	32
SCtsmuq	85.77	19	85.81	20
SCtsmtq	126.96	183	126.63	179
SCtsmsq	91.13	38	91.08	37
SCtsmtsq	107.93	92	107.04	91
SCtsmuy	84.05	15	81.67	11
SCtsmty	116.93	127	118.26	134
SCuqum	109.39	98	110.59	102
SCuqtm	132.75	206	131.52	202
SCuqsm	114.44	118	112.53	113
SCuqtsm	102.80	73	101.32	63
SCuquq	119.76	143	120.96	149
SCuqtq	138.01	217	136.01	216
SCuqsq	123.69	168	123.09	162
SCuqtsq	117.98	133	117.58	130
SCuquy	118.27	135	117.14	128
SCuqty	131.63	202	133.69	211
SCtqum	77.54	1	77.37	1
SCtqtm	133.12	208	130.79	199
SCtqsm	79.70	6	77.67	2
SCtqtsm	101.74	66	99.43	61
SCtquq	84.33	16	83.47	14
SCtqtq	135.63	216	134.85	214
SCtqsq	89.43	29	87.74	22
SCtqtsq	119.58	142	117.27	129
SCtquy	80.97	9	79.15	6
SCtqty	124.58	171	123.88	165
SCsqum	107.33	90	108.57	98
SCsqtm	133.62	210	132.99	209
SCsqsm	112.88	115	112.53	114

(Continued)

Table F-1 Continued

	MAD		RMSE	
	Mean rank	Rank	Mean rank	Rank
SCsqtsm	104.64	83	103.05	73
SCsquq	116.70	125	118.37	135
SCsqtq	138.67	219	137.70	219
SCsqsq	122.86	162	123.01	160
SCsqtsq	118.62	138	119.51	142
SCsquy	116.15	123	114.61	118
SCsqty	134.76	213	136.49	217
SCtsqum	80.95	8	81.61	10
SCtsqtm	127.46	186	126.85	181
SCtsqsm	83.85	14	83.74	16
SCtsqtsm	97.88	59	97.24	58
SCtsquq	89.08	27	89.16	29
SCtsqtq	130.95	199	130.51	197
SCtsqsq	94.74	52	93.99	50
SCtsqtsq	115.45	119	114.76	119
SCtsquy	86.63	21	84.21	18
SCtsqty	121.08	146	122.40	153
SCuyum	100.11	62	102.08	67
SCuytm	124.61	172	124.84	169
SCuysm	102.80	72	102.43	68
SCuytsm	95.76	56	95.60	56
SCuyuq	109.14	96	110.29	101
SCuytq	130.53	198	129.33	194
SCuysq	113.66	117	113.76	117
SCuytsq	110.39	101	109.33	99
SCuyuy	111.50	107	108.07	96
SCuyty	126.36	180	125.92	175
SCtyum	89.50	30	92.56	45
SCtytm	133.56	209	132.70	206
SCtysm	92.13	43	95.40	55
SCtytsm	104.39	82	104.54	81
SCtyuq	95.07	53	98.77	59
SCtytq	138.55	218	136.93	218
SCtysq	103.17	75	106.54	88
SCtytsq	121.12	147	122.66	156
SCtyuy	94.02	49	95.31	54
SCtyty	129.56	196	133.70	212
WCumum	102.82	74	103.28	74
WCumtm	119.24	141	119.20	139
WCumsm	106.01	87	105.35	84
WCumtsm	89.00	25	88.05	24
WCumuq	111.34	106	112.13	109
WCumtq	123.33	166	121.46	150
WCumsq	115.71	121	117.05	126
WCumtsq	117.64	132	117.06	127

(*Continued*)

Table F-1 Continued

	MAD		RMSE	
	Mean rank	Rank	Mean rank	Rank
WCumuy	104.20	80	103.78	76
WCumty	121.62	150	125.56	173
WCtmum	82.17	11	83.67	15
WCtmtm	131.87	204	130.88	200
WCtmsm	87.41	22	87.93	23
WCtmtsm	93.85	46	92.41	44
WCtmuq	91.31	40	91.99	39
WCtmtq	134.12	211	132.75	207
WCtmsq	100.33	63	101.88	65
WCtmtsq	125.66	176	124.17	167
WCtmuy	90.21	34	89.57	31
WCtmty	128.24	191	131.91	205
WCsmum	102.63	70	102.91	72
WCsmtm	117.13	129	117.83	132
WCsmsm	108.46	94	107.19	93
WCsmtsm	91.18	39	90.77	36
WCsmuq	112.15	114	112.41	112
WCsmtq	122.68	160	122.51	154
WCsmsq	118.46	136	119.13	138
WCsmtsq	118.72	139	118.17	133
WCsmuy	104.32	81	102.64	70
WCsmty	122.26	156	125.60	174
WCtsmum	82.52	12	82.93	13
WCtsmtm	123.43	167	122.74	158
WCtsmsm	90.72	36	89.77	33
WCtsmtsm	91.71	41	89.98	34
WCtsmuq	91.99	42	92.32	42
WCtsmtq	127.58	188	126.28	176
WCtsmsq	102.00	67	102.81	71
WCtsmtsq	121.66	152	120.82	148
WCtsmuy	89.65	31	88.17	25
WCtsmty	122.06	154	126.44	178
WCuqum	105.88	85	106.73	89
WCuqtm	120.26	145	120.48	146
WCuqsm	111.59	109	110.93	104
WCuqtsm	93.12	44	92.39	43
WCuquq	119.00	140	119.46	141
WCuqtq	126.67	181	124.73	168
WCuqsq	122.85	161	123.41	164
WCuqtsq	120.02	144	119.87	143
WCuquy	112.07	112	109.72	100
WCuqty	124.83	173	127.98	188
WCtqum	79.64	5	79.73	8
WCtqtm	131.10	200	129.16	193
WCtqsm	85.78	20	85.88	21
WCtqtsm	94.68	51	93.20	46
WCtquq	90.18	33	91.14	38

(Continued)

Table F-1 Continued

	MAD		RMSE	
	Mean rank	Rank	Mean rank	Rank
WCtqtq	134.44	212	132.94	208
WCtqsq	99.61	61	100.72	62
WCtqtsq	126.21	178	124.87	170
WCtquy	88.52	23	88.56	28
WCtqty	128.23	190	131.76	203
WCsqum	104.78	84	105.90	85
WCsqtm	122.46	158	123.01	161
WCsqsm	111.93	111	111.79	108
WCsqtsm	96.15	57	95.20	52
WCsquq	117.07	128	117.83	131
WCsqtq	128.71	194	127.64	187
WCsqsq	123.13	164	123.95	166
WCsqtsq	121.64	151	122.10	152
WCsquy	108.23	93	106.18	87
WCsqty	124.17	170	127.39	185
WCtsqum	83.39	13	82.83	12
WCtsqtm	128.61	193	126.43	177
WCtsqsm	88.55	24	88.28	27
WCtsqtsm	95.13	54	93.79	48
WCtsquq	93.99	48	93.71	47
WCtsqtq	132.83	207	130.63	198
WCtsqsq	102.76	71	103.36	75
WCtsqtsq	125.43	175	124.99	171
WCtsquy	90.10	32	89.26	30
WCtsqty	124.03	169	128.21	189
WCuyum	101.61	65	101.98	66
WCuytm	121.23	149	122.73	157
WCuysm	102.41	68	101.87	64
WCuytsm	89.11	28	90.40	35
WCuyuq	111.79	110	111.43	107
WCuytq	127.78	189	127.52	186
WCuysq	113.30	116	114.89	120
WCuytsq	117.35	130	118.67	137
WCuyuy	111.14	105	107.31	94
WCuyty	116.75	126	118.52	136
WCtyum	93.90	47	95.21	53
WCtytm	134.82	214	135.83	215
WCtysm	95.34	55	95.90	57
WCtytsm	100.36	64	102.59	69
WCtyuq	102.61	69	103.84	77
WCtytq	141.19	220	140.98	220
WCtysq	109.18	97	110.67	103
WCtytsq	127.35	185	128.90	191
WCtyuy	93.39	45	92.15	40
WCtyty	129.03	195	133.01	210

Appendix G - Hierarchical Forecasting Simulation Results Comparisons (2005 ~ 2007)

Table G-6 Hierarchical forecasting simulation results

	MAD		RMSE		Total inventory costs				
	Mean rank[6]	Rank[7]	Mean rank[6]	Rank[7]	Costs	HF-*um*[8]	Rank[7]	LN (HF/*tsm*)[9]	Rank[7]
TD1um	115.80	122	116.08	124	₩1,297,371,653	₩518,429,400	205	96.84	214
TD1tm	121.82	153	120.45	145	₩1,284,977,980	₩506,035,727	199	21.84	64
TD1sm	115.49	120	115.57	121	₩1,498,622,344	₩719,680,091	218	107.39	218
TD1tsm	91.06	37	92.17	41	₩1,201,187,680	₩422,245,427	191	6.04	30
TD1uq	129.67	197	129.06	192	₩1,289,275,626	₩510,333,373	201	126.76	219
TD1tq	127.29	184	125.43	172	₩1,216,999,330	₩438,057,077	195	19.56	57
TD1sq	127.54	187	128.66	190	₩1,818,557,369	₩1,039,615,116	220	129.25	220
TD1tsq	110.91	103	111.33	106	₩1,598,703,780	₩819,761,527	219	23.86	71
TD1uy	123.18	165	120.63	147	₩1,276,903,532	₩497,961,279	198	103.50	217
TD1ty	126.89	182	126.70	180	₩1,164,660,122	₩385,717,869	175	42.59	121
TD2um	103.90	79	105.10	82	₩771,715,898	-₩7,226,355	29	53.54	142
TD2tm	118.60	137	116.04	123	₩854,063,515	₩75,121,262	109	-0.98	19
TD2sm	106.06	88	106.00	86	₩912,760,261	₩133,818,008	137	61.51	156
TD2tsm	84.90	17	85.26	19	₩770,753,404	-₩8,188,848	27	-25.38	1
TD2uq	118.15	134	119.38	140	₩787,978,172	₩9,035,919	52	78.55	189
TD2tq	122.44	157	120.02	144	₩799,522,960	₩20,580,707	65	-3.67	16
TD2sq	121.16	148	122.07	151	₩1,123,507,820	₩344,565,567	170	83.34	202
TD2tsq	106.54	89	107.18	92	₩1,050,381,845	₩271,439,592	163	2.15	23
TD2uy	107.56	91	106.75	90	₩774,645,225	-₩4,297,028	33	55.99	150
TD2ty	122.07	155	123.27	163	₩786,891,652	₩7,949,399	49	16.76	48
SCumum	103.23	76	104.37	80	₩780,851,097	₩1,908,844	39	54.29	145
SCumtm	126.30	179	126.86	182	₩1,205,874,413	₩426,932,161	192	72.75	179
SCumsm	103.38	77	104.32	79	₩841,595,350	₩62,653,097	100	64.10	160
SCumtsm	94.06	50	93.83	49	₩863,951,076	₩85,008,823	116	32.10	94
SCumuq	110.56	102	112.15	110	₩811,323,226	₩32,380,973	78	69.83	174
SCumtq	131.78	203	130.02	196	₩1,156,796,378	₩377,854,125	173	84.34	206
SCumsq	112.12	113	112.77	116	₩904,806,802	₩125,864,549	133	72.12	177
SCumtsq	110.35	100	110.98	105	₩943,474,150	₩164,531,897	150	53.93	144
SCumuy	108.52	95	107.62	95	₩1,161,312,366	₩382,370,114	174	82.48	200
SCumty	125.39	174	127.15	184	₩1,289,852,844	₩510,910,591	202	84.29	204
SCtmum	79.34	4	79.21	7	₩726,002,802	-₩52,939,451	7	-18.42	3
SCtmtm	132.75	205	131.76	204	₩820,724,706	₩41,782,454	85	22.73	67

(Continued)

[6] Mean ranks for the forecasting method in terms of MAD or RMSE over the 300 items;

[7] The rank of the forecasting method among the 220 hierarchical forecasting methods in terms of the criterion in the left column;

[8] The total inventory costs of each hierarchical forecasting method deducted by the total inventory costs of *um*; and

[9] The sum of natural log relative error [$\ln(\text{error}_{HF}/\text{error}_{tsm})$] for the total inventory costs of each forecasting method over the 300 items.

Table G-1 Continued

	MAD		RMSE		Total inventory costs				
	Mean rank	Rank	Mean rank	Rank	Costs	HF-*um*	Rank	LN (HF/*tsm*)	Rank
SCtmsm	79.05	3	78.38	3	₩730,726,074	-₩48,216,179	10	-11.79	8
SCtmtsm	99.25	60	98.78	60	₩803,679,850	₩24,737,597	72	-13.54	6
SCtmuq	84.92	18	83.93	17	₩740,191,882	-₩38,750,371	12	-2.66	17
SCtmtq	134.88	215	133.93	213	₩839,930,034	₩60,987,782	98	23.00	69
SCtmsq	89.02	26	88.25	26	₩795,862,018	₩16,919,765	59	6.00	29
SCtmtsq	117.59	131	116.14	125	₩872,434,952	₩93,492,699	122	15.65	46
SCtmuy	80.01	7	79.00	4	₩1,027,042,295	₩248,100,042	161	4.59	26
SCtmty	123.00	163	122.80	159	₩1,079,385,032	₩300,442,779	167	35.78	108
SCsmum	103.75	78	104.18	78	₩828,519,164	₩49,576,911	91	56.05	151
SCsmtm	125.83	177	127.04	183	₩1,207,929,635	₩428,987,383	193	74.11	183
SCsmsm	105.99	86	105.15	83	₩885,115,763	₩106,173,510	126	64.18	161
SCsmtsm	96.24	58	94.84	51	₩872,116,714	₩93,174,461	121	31.72	93
SCsmuq	111.58	108	112.32	111	₩858,107,344	₩79,165,091	111	69.39	171
SCsmtq	131.51	201	131.50	201	₩1,186,685,985	₩407,743,732	189	83.48	203
SCsmsq	116.25	124	116.02	122	₩929,966,505	₩151,024,252	146	72.21	178
SCsmtsq	111.01	104	112.69	115	₩962,553,907	₩183,611,654	154	52.82	140
SCsmuy	110.06	99	108.08	97	₩1,211,488,999	₩432,546,746	194	81.98	199
SCsmty	128.55	192	129.57	195	₩1,319,877,599	₩540,935,347	210	81.98	198
SCtsmum	78.70	2	79.07	5	₩730,245,647	-₩48,696,606	9	1.29	21
SCtsmtm	122.57	159	122.53	155	₩901,933,968	₩122,991,715	132	29.77	90
SCtsmsm	81.30	10	81.03	9	₩748,440,413	-₩30,501,840	16	9.99	33
SCtsmtsm	90.71	35	89.68	32	₩780,007,802	₩1,065,549	37	-8.28	11
SCtsmuq	85.77	19	85.81	20	₩748,193,308	-₩30,748,944	15	14.62	41
SCtsmtq	126.96	183	126.63	179	₩909,257,343	₩130,315,091	134	36.26	109
SCtsmsq	91.13	38	91.08	37	₩811,127,257	₩32,185,005	77	20.95	62
SCtsmtsq	107.93	92	107.04	91	₩853,436,369	₩74,494,116	108	11.87	36
SCtsmuy	84.05	15	81.67	11	₩1,059,065,064	₩280,122,811	165	26.55	79
SCtsmty	116.93	127	118.26	134	₩1,097,099,010	₩318,156,757	168	40.72	117
SCuqum	109.39	98	110.59	102	₩791,577,425	₩12,635,172	57	68.86	170
SCuqtm	132.75	206	131.52	202	₩1,226,847,655	₩447,905,402	196	87.30	207
SCuqsm	114.44	118	112.53	113	₩851,688,467	₩72,746,214	106	75.14	185
SCuqtsm	102.80	73	101.32	63	₩878,725,716	₩99,783,463	123	41.34	118
SCuquq	119.76	143	120.96	149	₩819,859,183	₩40,916,930	84	80.14	195
SCuqtq	138.01	217	136.01	216	₩1,172,274,765	₩393,332,512	180	97.91	215
SCuqsq	123.69	168	123.09	162	₩915,212,040	₩136,269,788	139	82.62	201
SCuqtsq	117.98	133	117.58	130	₩956,899,709	₩177,957,456	152	65.65	165
SCuquy	118.27	135	117.14	128	₩1,182,430,883	₩403,488,630	188	92.95	211
SCuqty	131.63	202	133.69	211	₩1,296,394,155	₩517,451,902	204	95.24	212
SCtqum	77.54	1	77.37	1	₩693,601,747	-₩85,340,505	1	-24.64	2
SCtqtm	133.12	208	130.79	199	₩798,643,689	₩19,701,436	64	17.18	51
SCtqsm	79.70	6	77.67	2	₩702,219,843	-₩76,722,409	3	-16.08	5
SCtqtsm	101.74	66	99.43	61	₩773,325,925	-₩5,616,328	30	-17.04	4
SCtquq	84.33	16	83.47	14	₩706,669,261	-₩72,272,992	4	-9.04	10
SCtqtq	135.63	216	134.85	214	₩814,055,960	₩35,113,707	80	15.84	47
SCtqsq	89.43	29	87.74	22	₩771,252,823	-₩7,689,430	28	0.96	20
SCtqtsq	119.58	142	117.27	129	₩846,733,720	₩67,791,467	102	12.25	37
SCtquy	80.97	9	79.15	6	₩991,497,949	₩212,555,696	158	-1.14	18

(Continued)

Table G-1 Continued

	MAD		RMSE		Total inventory costs				
	Mean rank	Rank	Mean rank	Rank	Costs	HF-*um*	Rank	LN (HF/*tsm*)	Rank
SCtqty	124.58	171	123.88	165	₩1,053,941,420	₩274,999,167	164	32.13	95
SCsqum	107.33	90	108.57	98	₩940,265,101	₩161,322,848	148	65.46	163
SCsqtm	133.62	210	132.99	209	₩1,317,188,854	₩538,246,601	209	88.00	209
SCsqsm	112.88	115	112.53	114	₩942,413,391	₩163,471,138	149	71.57	176
SCsqtsm	104.64	83	103.05	73	₩932,183,368	₩153,241,115	147	42.29	120
SCsquq	116.70	125	118.37	135	₩962,029,570	₩183,087,317	153	76.17	187
SCsqtq	138.67	219	137.70	219	₩1,285,423,294	₩506,481,042	200	98.18	216
SCsqsq	122.86	162	123.01	160	₩980,153,669	₩201,211,416	157	80.00	194
SCsqtsq	118.62	138	119.51	142	₩1,050,001,366	₩271,059,113	162	68.35	168
SCsquy	116.15	123	114.61	118	₩1,327,825,555	₩548,883,302	211	90.29	210
SCsqty	134.76	213	136.49	217	₩1,432,940,798	₩653,998,545	217	96.81	213
SCtsqum	80.95	8	81.61	10	₩833,085,383	₩54,143,131	92	-3.90	15
SCtsqtm	127.46	186	126.85	181	₩910,049,252	₩131,106,999	135	29.63	89
SCtsqsm	83.85	14	83.74	16	₩786,260,960	₩7,318,707	47	4.41	25
SCtsqtsm	97.88	59	97.24	58	₩819,406,914	₩40,464,661	82	-7.31	12
SCtsquq	89.08	27	89.16	29	₩848,480,754	₩69,538,502	104	9.71	32
SCtsqtq	130.95	199	130.51	197	₩928,240,757	₩149,298,504	145	29.27	87
SCtsqsq	94.74	52	93.99	50	₩839,873,763	₩60,931,510	97	17.87	54
SCtsqtsq	115.45	119	114.76	119	₩913,666,873	₩134,724,621	138	17.41	52
SCtsquy	86.63	21	84.21	18	₩1,143,739,608	₩364,797,355	172	17.70	53
SCtsqty	121.08	146	122.40	153	₩1,189,443,687	₩410,501,434	190	39.51	114
SCuyum	100.11	62	102.08	67	₩763,061,418	-₩15,880,835	23	53.57	143
SCuytm	124.61	172	124.84	169	₩1,178,317,100	₩399,374,847	184	65.70	166
SCuysm	102.80	72	102.43	68	₩798,553,013	₩19,610,760	63	55.61	149
SCuytsm	95.76	56	95.60	56	₩850,360,325	₩71,418,072	105	29.62	88
SCuyuq	109.14	96	110.29	101	₩781,629,520	₩2,687,267	42	63.30	159
SCuytq	130.53	198	129.33	194	₩1,101,853,761	₩322,911,509	169	74.37	184
SCuysq	113.66	117	113.76	117	₩864,855,028	₩85,912,775	117	65.94	167
SCuytsq	110.39	101	109.33	99	₩920,408,032	₩141,465,780	141	44.27	126
SCuyuy	111.50	107	108.07	96	₩1,060,238,351	₩281,296,098	166	78.75	190
SCuyty	126.36	180	125.92	175	₩1,178,892,113	₩399,949,860	185	73.44	182
SCtyum	89.50	30	92.56	45	₩702,216,842	-₩76,725,411	2	13.53	39
SCtytm	133.56	209	132.70	206	₩794,344,921	₩15,402,668	58	38.56	111
SCtysm	92.13	43	95.40	55	₩715,748,865	-₩63,193,388	6	19.80	58
SCtytsm	104.39	82	104.54	81	₩762,744,963	-₩16,197,290	22	4.32	24
SCtyuq	95.07	53	98.77	59	₩714,672,940	-₩64,269,313	5	22.51	65
SCtytq	138.55	218	136.93	218	₩801,943,719	₩23,001,466	68	42.26	119
SCtysq	103.17	75	106.54	88	₩784,043,581	₩5,101,328	45	30.80	92
SCtytsq	121.12	147	122.66	156	₩823,139,479	₩44,197,226	88	24.85	74
SCtyuy	94.02	49	95.31	54	₩993,229,897	₩214,287,644	159	38.75	112
SCtyty	129.56	196	133.70	212	₩1,007,623,544	₩228,681,291	160	45.04	128
WCumum	102.82	74	103.28	74	₩779,073,203	₩130,950	36	52.41	138
WCumtm	119.24	141	119.20	139	₩773,869,906	-₩5,072,347	32	21.56	63
WCumsm	106.01	87	105.35	84	₩840,466,528	₩61,524,275	99	65.17	162
WCumtsm	89.00	25	88.05	24	₩758,279,736	-₩20,662,517	18	10.45	34
WCumuq	111.34	106	112.13	109	₩807,528,276	₩28,586,024	74	69.76	173
WCumtq	123.33	166	121.46	150	₩761,152,212	-₩17,790,041	20	26.52	78

(*Continued*)

Table G-1 Continued

	MAD		RMSE		Total inventory costs				
	Mean rank	Rank	Mean rank	Rank	Costs	HF-*um*	Rank	LN (HF/*tsm*)	Rank
WCumsq	115.71	121	117.05	126	₩886,940,427	₩107,998,175	128	72.89	180
WCumtsq	117.64	132	117.06	127	₩861,553,487	₩82,611,234	114	30.75	91
WCumuy	104.20	80	103.78	76	₩1,337,962,681	₩559,020,428	214	81.93	197
WCumty	121.62	150	125.56	173	₩1,169,793,269	₩390,851,016	177	55.60	147
WCtmum	82.17	11	83.67	15	₩756,997,140	-₩21,945,113	17	14.70	42
WCtmtm	131.87	204	130.88	200	₩817,146,107	₩38,203,855	81	24.40	72
WCtmsm	87.41	22	87.93	23	₩826,137,247	₩47,194,994	90	34.74	104
WCtmtsm	93.85	46	92.41	44	₩796,912,497	₩17,970,244	60	-6.91	13
WCtmuq	91.31	40	91.99	39	₩785,064,652	₩6,122,399	46	32.65	98
WCtmtq	134.12	211	132.75	207	₩824,868,592	₩45,926,339	89	23.24	70
WCtmsq	100.33	63	101.88	65	₩885,426,849	₩106,484,597	127	47.50	129
WCtmtsq	125.66	176	124.17	167	₩925,407,367	₩146,465,115	144	27.36	84
WCtmuy	90.21	34	89.57	31	₩1,309,566,180	₩530,623,927	208	55.30	146
WCtmty	128.24	191	131.91	205	₩1,173,789,234	₩394,846,981	181	51.82	136
WCsmum	102.63	70	102.91	72	₩802,080,605	₩23,138,352	69	53.38	141
WCsmtm	117.13	129	117.83	132	₩799,863,630	₩20,921,377	66	26.78	80
WCsmsm	108.46	94	107.19	93	₩881,604,720	₩102,662,467	124	65.52	164
WCsmtsm	91.18	39	90.77	36	₩790,889,818	₩11,947,565	56	12.92	38
WCsmuq	112.15	114	112.41	112	₩835,221,735	₩56,279,482	94	69.96	175
WCsmtq	122.68	160	122.51	154	₩790,576,879	₩11,634,626	55	32.64	97
WCsmsq	118.46	136	119.13	138	₩924,376,969	₩145,434,716	143	73.16	181
WCsmtsq	118.72	139	118.17	133	₩893,890,407	₩114,948,154	129	33.06	100
WCsmuy	104.32	81	102.64	70	₩1,335,436,872	₩556,494,619	213	79.53	192
WCsmty	122.26	156	125.60	174	₩1,176,265,315	₩397,323,062	183	55.60	148
WCtsmum	82.52	12	82.93	13	₩747,740,810	-₩31,201,442	14	15.07	45
WCtsmtm	123.43	167	122.74	158	₩804,133,818	₩25,191,565	73	16.86	50
WCtsmsm	90.72	36	89.77	33	₩822,160,863	₩43,218,610	87	35.77	107
WCtsmtsm	91.71	41	89.98	34	₩787,884,582	₩8,942,330	51	-6.33	14
WCtsmuq	91.99	42	92.32	42	₩776,019,061	-₩2,923,192	34	33.28	101
WCtsmtq	127.58	188	126.28	176	₩802,739,066	₩23,796,814	71	18.28	55
WCtsmsq	102.00	67	102.81	71	₩869,166,363	₩90,224,110	119	44.94	127
WCtsmtsq	121.66	152	120.82	148	₩901,298,819	₩122,356,566	131	22.63	66
WCtsmuy	89.65	31	88.17	25	₩1,304,531,909	₩525,589,656	206	52.01	137
WCtsmty	122.06	154	126.44	178	₩1,172,030,145	₩393,087,892	179	48.62	134
WCuqum	105.88	85	106.73	89	₩781,156,444	₩2,214,191	41	59.65	155
WCuqtm	120.26	145	120.48	146	₩778,328,564	-₩613,689	35	28.18	86
WCuqsm	111.59	109	110.93	104	₩844,150,886	₩65,208,633	101	69.53	172
WCuqtsm	93.12	44	92.39	43	₩761,937,401	-₩17,004,852	21	14.99	44
WCuquq	119.00	140	119.46	141	₩812,202,332	₩33,260,079	79	78.30	188
WCuqtq	126.67	181	124.73	168	₩765,910,726	-₩13,031,526	25	33.72	103
WCuqsq	122.85	161	123.41	164	₩896,428,526	₩117,486,273	130	78.96	191
WCuqtsq	120.02	144	119.87	143	₩867,243,218	₩88,300,965	118	34.77	105
WCuquy	112.07	112	109.72	100	₩1,344,361,608	₩565,419,355	215	87.31	208
WCuqty	124.83	173	127.98	188	₩1,171,479,891	₩392,537,639	178	59.52	154
WCtqum	79.64	5	79.73	8	₩728,256,615	-₩50,685,638	8	5.29	27
WCtqtm	131.10	200	129.16	193	₩798,231,792	₩19,289,539	62	14.73	43
WCtqsm	85.78	20	85.88	21	₩802,284,947	₩23,342,694	70	25.85	76

(*Continued*)

Table G-1 Continued

	MAD		RMSE		Total inventory costs				
	Mean rank	Rank	Mean rank	Rank	Costs	HF-*um*	Rank	LN (HF/*tsm*)	Rank
WCtqtsm	94.68	51	93.20	46	₩782,633,097	₩3,690,845	43	-12.68	7
WCtquq	90.18	33	91.14	38	₩759,015,207	-₩19,927,046	19	24.45	73
WCtqtq	134.44	212	132.94	208	₩808,126,980	₩29,184,728	75	14.46	40
WCtqsq	99.61	61	100.72	62	₩863,562,880	₩84,620,627	115	38.90	113
WCtqtsq	126.21	178	124.87	170	₩912,739,321	₩133,797,068	136	22.85	68
WCtquy	88.52	23	88.56	28	₩1,293,165,408	₩514,223,156	203	47.76	132
WCtqty	128.23	190	131.76	203	₩1,169,187,203	₩390,244,951	176	48.48	133
WCsqum	104.78	84	105.90	85	₩848,298,802	₩69,356,550	103	59.10	152
WCsqtm	122.46	158	123.01	161	₩860,676,055	₩81,733,802	113	37.20	110
WCsqsm	111.93	111	111.79	108	₩921,403,054	₩142,460,801	142	68.40	169
WCsqtsm	96.15	57	95.20	52	₩836,659,068	₩57,716,816	95	20.32	60
WCsquq	117.07	128	117.83	131	₩884,685,348	₩105,743,096	125	76.01	186
WCsqtq	128.71	194	127.64	187	₩852,106,918	₩73,164,665	107	43.44	123
WCsqsq	123.13	164	123.95	166	₩977,070,648	₩198,128,395	156	79.85	193
WCsqtsq	121.64	151	122.10	152	₩949,170,448	₩170,228,195	151	39.59	115
WCsquy	108.23	93	106.18	87	₩1,355,332,352	₩576,390,099	216	81.92	196
WCsqty	124.17	170	127.39	185	₩1,181,394,664	₩402,452,412	186	59.13	153
WCtsqum	83.39	13	82.83	12	₩783,713,973	₩4,771,720	44	9.08	31
WCtsqtm	128.61	193	126.43	177	₩859,385,696	₩80,443,443	112	19.00	56
WCtsqsm	88.55	24	88.28	27	₩854,524,732	₩75,582,479	110	28.15	85
WCtsqtsm	95.13	54	93.79	48	₩836,978,887	₩58,036,634	96	-9.39	9
WCtsquq	93.99	48	93.71	47	₩819,425,092	₩40,482,839	83	26.93	82
WCtsqtq	132.83	207	130.63	198	₩870,907,524	₩91,965,271	120	20.01	59
WCtsqsq	102.76	71	103.36	75	₩916,446,088	₩137,503,835	140	40.69	116
WCtsqtsq	125.43	175	124.99	171	₩972,117,307	₩193,175,054	155	25.46	75
WCtsquy	90.10	32	89.26	30	₩1,306,456,688	₩527,514,435	207	43.81	124
WCtsqty	124.03	169	128.21	189	₩1,174,526,160	₩395,583,907	182	47.57	130
WCuyum	101.61	65	101.98	66	₩767,420,877	-₩11,521,375	26	49.85	135
WCuytm	121.23	149	122.73	157	₩787,477,282	₩8,535,029	50	26.89	81
WCuysm	102.41	68	101.87	64	₩773,371,356	-₩5,570,897	31	52.50	139
WCuytsm	89.11	28	90.40	35	₩745,773,185	-₩33,169,068	13	11.02	35
WCuyuq	111.79	110	111.43	107	₩789,760,473	₩10,818,220	54	62.66	158
WCuytq	127.78	189	127.52	186	₩781,151,346	₩2,209,093	40	33.40	102
WCuysq	113.30	116	114.89	120	₩800,631,743	₩21,689,490	67	61.93	157
WCuytsq	117.35	130	118.67	137	₩786,519,567	₩7,577,314	48	26.18	77
WCuyuy	111.14	105	107.31	94	₩1,333,774,117	₩554,831,865	212	84.30	205
WCuyty	116.75	126	118.52	136	₩1,126,777,335	₩347,835,082	171	42.61	122
WCtyum	93.90	47	95.21	53	₩739,345,856	-₩39,596,397	11	5.38	28
WCtytm	134.82	214	135.83	215	₩809,566,242	₩30,623,989	76	35.00	106
WCtysm	95.34	55	95.90	57	₩780,231,795	₩1,289,543	38	16.78	49
WCtytsm	100.36	64	102.59	69	₩789,731,749	₩10,789,497	53	1.95	22
WCtyuq	102.61	69	103.84	77	₩764,988,840	-₩13,953,412	24	20.66	61
WCtytq	141.19	220	140.98	220	₩820,906,630	₩41,964,377	86	43.89	125
WCtysq	109.18	97	110.67	103	₩797,394,222	₩18,451,969	61	32.63	96
WCtytsq	127.35	185	128.90	191	₩834,727,752	₩55,785,499	93	27.07	83
WCtyuy	93.39	45	92.15	40	₩1,270,529,795	₩491,587,542	197	32.98	99
WCtyty	129.03	195	133.01	210	₩1,182,294,273	₩403,352,021	187	47.72	131

References

Agripino, M., Cathcart, T. and Mathaisel, D. (2002) 'A Lean sustainment enterprise model for military systems', *Acquisition Review Quarterly*, pp. 274 - 296.

Altman, D. G. and Royston, P. (2000) 'What do we mean by validating a prognostic model?', *Statistics in Medicine,* 19, (4), pp. 453 - 473.

An, B.-K., Kim, T.-H. and Moon, S.-A. (2002) 'A study on the effect of the inventory policy on military supply chain performance: focused on system dynamics', *Military Operations Research Society of Korea,* 28, (2), pp. 1 - 19.

Armstrong, J. S. and Collopy, F. (1992) 'Error measures for generalizing about forecasting methods: empirical comparisons', *International Journal of Forecasting,* 8, (1), pp. 69 - 80.

Bacharach, S. B. (1989) 'Organisational theories: some criteria for evaluation', *Academy of Management Review,* 14, (4), pp. 496 - 515.

Balakrishnan, N. and Nevzorov, V. B. (2003) *A Primer on Statistical Distribution.* John Wiley & Sons: NJ.

Bertrand, J. W. M. and Fransoo, J. C. (2002) 'Operations management research methodologies using quantitative modeling', *Operations Management Research,* 22, (2), pp. 241 - 264.

Bowerman, B. L., O'Connell, R. T. and B., K. A. (2005) *Forecasting, Time Series, and Regression, 4th edn.* Thomson Brooks/Cole: Belmont.

Bowersox, D. J. and Closs, D. J. (1996) *Logistical Management: The Integrated Supply Chain Process.* McGraw-Hill.: London.

Box, G. E. P. and Cox, D. R. (1964) 'An analysis of transformations', *Journal of the Royal Statistical Society. Series B,* 26, (2), pp. 211 - 252.

Boylan, J. E., Syntetos, A. A. and Karakostas, G. C. (2008) 'Classification for forecasting and stock control: a case study', *Journal of Operational Research Society,* 59, pp. 473 - 481.

Brier, G. W. (1950) 'Verification of forecasts expressed in terms of probability', *Monthly Weather Review,* 78, (1), pp. 1 - 3.

Businger, M. P. and Read, R. R. (1999) 'Identification of demand patterns for selective processing: a case study', *International Journal of Management Science,* 27, pp. 189 - 200.

Chatfield, C. (2004) *The Analysis of Time Series: an Introduction, 6th edn.* Chapman & Hall/CRC: London.

Chen, H., Li, L., Wang, J. and Xu, Q. (2005) 'Military supply chain integration', *Services Systems and Services Management, Proceedings of ICSSSM '05,* 1, pp. 674 - 679

Choi, S., Jang, J., Jang, K. and Seon, M. (2005) *A research about optimal spare parts demand assessment during peace time.* Seoul, South Korea: Korea Institute for Defence Analysis

Conger, A. J. (1974) 'A revised definition for suppressor variables: A guide to their identification and interpretation', *Educational and Psychological Measurement,* 34, pp. 35 - 46.

Cook, D. and Weisberg, S. (1982) *Residuals and Influence in Regression.* Chapman and Hall: NY.

Croston, J. D. (1972) 'Forecasting and stock control for intermittent demands', *Operational Research Quarterly* 23, (3), pp. 289-303.

Cryer, J. D. and Chan, K.-S. (2008) *Time Series Analysis with Application in R.* Springer: NY.

Dangerfield, B. J. and Morris, J. S. (1992) 'Top-down or bottom-up: aggregate versus disaggregate extrapolations', *International Journal of Forecasting*, 8, (2), pp. 233 - 241.

Defence Industry Daily (2009) *Korea's KDX-III AEGIS Destroyers*. Available at: http://www.defenseindustrydaily.com/drs-wins-multiplexing-contract-for-korean-aegis-destroyers-0431/ (Accessed: 1 December 2009).

Dekker, M., Donselaar. K. v. and Ouwehand, a. P. (2004) 'How to use aggregation and combined forecasting to improve seasonal demand forecasts', *International Journal of Production Economics*, 90, (2), pp. 151 - 167.

DeLurgio, S. A. (1998) *Forecasting Principles and Applications*. Irwin McGraw-Hill: NY.

Deshpande, V., Cohen. M. A. and Donohue, K. (2003) 'An empirical study of service differentiation for weapon system service parts', *Operations Research*, 51, (4), pp. 518 - 530.

Dreiseitl, S. and Ohno-Machado, L. (2002) 'Logistic regression and artificial neural network classification models: a methodology review', *Journal of Biomedical Informatics*, 35, (5/6), pp. 352 - 359.

du Preez, J. and Witt, S. F. (2003) 'Univariate versus multivariate time series forecasting: an application to international tourism demand', *International Journal of Forecasting*, 19, (3), pp. 435 - 451.

Eaves, A. H. C. (2002) *Forecasting for the Ordering and Stock-holding of Consumable Spare Parts*. PhD Thesis. University of Lancaster.

Eaves, A. H. C. and Kingsman, B. G. (2004) 'Forecasting for the ordering and stock-holding of spare parts', *Journal of the Operational Research Society*, 55, pp. 431 - 247.

Eisenhardt, K. M. (1989) 'Building theories from case study research', *Academy of Management Review*, 14, (4), pp. 532 - 550.

Field, A. (2009) *Discovering Statistics Using SPSS, 3rd edn*. SAGE: London.

Fisher, R. A. (1922) 'On the interpretation of chi-square from contingency tables, and the calculation of P', *Journal of the Royal Statistical Society*, 85, (1), pp. 87 - 94.

Fliedner, E. B. and Lawrence, B. (1995) 'Forecasting system parent group formation: An empirical application of cluster analysis', *Journal of Operations Management*, 12, (2), pp. 119 - 130.

Fliedner, E. B. and Mabert, V. A. (1992) 'Constrained forecasting: some implementation guidelines', *Decision Sciences*, 23, (5), pp. 1143 - 1161.

Fliedner, G. (1999) 'An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation', *Computers & Operations Research*, 26, (10 -11), pp. 1133 - 1149.

Fliedner, G. (2001) 'Hierarchical forecasting: issues and use guidelines ', *Industrial Management & Data Systems*, 101, (1), pp. 5 - 12.

Fowler, A. (2003) 'The Royal Navy: JIT based ship spares provisioning ', in Johnston, R., Chambers, S., Harland, C., Harrison, A. and Slack, N.(eds) *Cases in Operations Management*. Prentice Hall: London, pp. 300 - 307.

Friedman, M. (1937) 'The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance ', *Journal of the American Statistical Association*, 32, pp. 675 - 701.

Gardner, E. S. (1985) 'Exponential smoothing: the state of the art', *Journal of Forecasting*, 4, pp. 1 - 28.

Gardner Jr., E. S. and Diaz-Saiz, J. (2002) 'Seasonal adjustment of inventory demand series: a case study', *International Journal of Forecasting*, 18, (1), pp. 117 - 223.

Ghobbar, A. A. and Friend, C. H. (2002) 'Sources of intermittent demand for aircraft spare parts within airline operations', *Journal of Air Transport Management*, 8, pp. 221 - 231.

Ghobbar, A. A. and Friend, C. H. (2003) 'Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model', *Computers & Operations Research*, 30, (14), pp. 2097 - 2114.

Gross, C. W. and Sohl, J. E. (1990) 'Disaggregation methods to expedite product line forecasting', *Journal of Forecasting*, 9, pp. 233 - 254.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996) 'Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors', *Statistics in Medicine*, 15, (4), pp. 361 - 387.

Heuts, R. M. J., Strijbosch, L. W. G. and Schoot, E. H. M. v. d. (1999) 'A combined forecast-inventory control procedure for spare parts', *Flexible Automation & Intelligent Manufacturing*, pp. 1 - 15.

Hinton Jr., H. L. (1999) *Defense inventory, continuing challenges in managing inventories and avoiding adverse operational effects*. Washington, D.C.: US General Accounting Office

Hosmer, D. W. and Lemeshow, S. (2000) *Applied Logistic Regression, 2nd edn*. John Wiley & Sons: NY.

Howitt, D. and Cramer, D. (2008) *Introduction to Statistics in Psychology: 4th edn*. Pearson Education: Essex.

Hyndman, R. J. (2010) *Forecast 2.04*. Available at: http://www.stats.bris.ac.uk/R/ (Accessed: 17 April 2010).

Hyndman, R. J., Ahmed, R. A. and Athanasopoulos, G. (2007) 'Optimal combination forecasts for hierarchical time series ', *Monash University*, pp. 1 - 21.

Hyndman, R. J., Makridakis, S. and Wheelwright, S. C. (1998) *Forecasting: Methods and Applications, 3rd edn*. John Wiley & Sons: NY.

Jiafu, R., Zongfang, Z. and Fang, Z. (2009) 'The forecasting models for spare parts based on ARMA', *2009 World Congress on Computer Science and Information Engineering*, 4, pp. 499 - 503.

Jie-Ae, S. (2002) *South, North Korea clash at sea*. Available at: http://edition.cnn.com/2002/WORLD/asiapcf/east/06/29/korea.warships/ (Accessed: 13 November 2009).

Johnston, F. R. (1980) 'An interactive stock control system with a strategic management role', *Journal of Operational Research Society*, 31, pp. 1069 - 1084.

Johnston, F. R. and Boylan, J. E. (1996) 'Forecasting for items with intermittent demand', *Journal of Operational Research Society*, 47, pp. 113 - 121.

Kahn, K. B. (1998) 'Revisiting top-down versus bottom-up forecasting', *The Journal of Business Forecasting*, Summer, pp. 14 - 19.

Kanji, G. K. (2006) *100 Statistical Tests: 3rd edn*. SAGE Publications: London.

Kling, J. L. and Bessler, D. A. (1985) 'A comparison of multivariate forecasting procedures for economic time series', *International Journal of Forecasting*, 1, (1), pp. 5 - 24.

Korean Defence White Paper. (2008)

Korean Navy. (2003) *Korean Navy Regulation Rule Book*.

Krishnamoorthy, K. (2006) *Handbook of Statistical Distributions with Applications*. Chapman & Hall/CRC: Florida.

Krzanowski, W. J. (1998) *An Introduction to Statistical Modelling*. Arnold: London.

Lancaster, B. P. (1999) 'Defining and interpreting suppressor effects: advantages and limitations', *Annual Meeting of the Southwest Educational Research Association*, pp. 1 - 21.

Lee, H. L., Padmanabhan, V. and Whang, S. (1997) 'Information distortion in a supply chain: the bullwhip effect', *Management Science,* 43, (4), pp. 546 - 558.

Lee, S.-Y. (2007) 'Diagnosis of maintenance budget and future development strategy', *The Korean Journal of Defense Analysis,* 76, pp. 43 - 80.

Leonard-Barton, D. (1990) 'A dual methodology for case studies: synergistic use of a longitudinal single site with replicated multiple sites', 1, (3), pp. 248 - 266.

MacDonald, S. D. (1997) *Navy Inventory Management Decision-Making.* MSc Thesis. Naval Postgraduate School.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. (1982) 'The accuracy of extrapolation (time series) methods: results of a forecasting competition', *Journal of Forecasting,* 1, (2), pp. 111 - 153.

Makridakis, S. and Winkler, R. L. (1983) 'Averages of forecasts: some empirical results', *Management Science,* 29, (9), pp. 987-996.

Markland, R. E. (1970) 'A comparative study of demand forecasting techniques for military helicopter spare parts', *Naval Research Logistics Quarterly,* 17, (1), pp. 103 - 119.

McCleary, R. (1980) *Applied Time Series Analysis for the Social Sciences.* SAGE Publications: London.

McCutcheon, D. M. and Meredith, J. R. (1993) 'Conducting case study research in operations management', 11, (3), pp. 239 - 256

Menard, S. W. (2002) *Applied Logistic Regression Analysis, 2nd edn.* Sage: Thousand Oaks, CA.

Miles, J. and Shevlin, M. (2001) *Applying Regression and Correlation: a Guide for Students and Researchers.* Sage publications: London.

Miller, J. G., Berry, W. L. and Lai, C.-Y. F. (2007) 'A comparison of alternative forecasting strategies for multi-stage production-inventory systems', *Decision Sciences,* 7, (4), pp. 714 - 724.

Miller Jr., R. G. (1986) *Beyond ANOVA, Basics of Applied Statistics.* John Wiley & Sons, Inc.: NY.

Moore, D. S., McCabe, G. P. and Craig, B. A. (2009) *Introduction to the Practice of Statistics: Sixth edn.* W. H. Freeman and Company: NY.

Morris, R. H. and Glassey, C. R. (1962) 'The dynamics and statistics of exponential smoothing operators', *Operations Research,* 11, (4), pp. 561 - 569.

Myers, R. H. (1990) *Classical and Modern Regression with Applications,2nd edn.* Duxbury: Boston, MA.

Narasimhan, S. L., McLeavey, D. W. and Billington, P. J. (1998) *Production Planning and Inventory Control, 2nd edn.* Prentice Hall: London.

Nelson, H. L. and Granger, C. W. J. (1979) 'Experience with using the Box-Cox transformation when forecasting economic time series ', *Journal of Econometrics,* 10, (1), pp. 57 - 69

Neter, J., Wasserman, W. and Kutner, M. H. (1990) *Applied Linear Statistical Models, 3rd edn.* Richard D. Irwin Inc.: Boston, MA.

Ottenbacher, K. J., Smith, P. M., Illig, S. B., Linn, R. T., Fiedler, R. C. and Granger, C. V. (2001) 'Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke', *Journal of Clinical Epidemiology,* 54, (11), pp. 1159 - 1165.

Pallant, J. (2005) *SPSS Survival Manual, 2nd edn.* Open University Press: Berkshire, UK.

Perlich, C., Provost, F. and Simonoff, J. S. (2003) 'Tree induction vs. logistic regression: a learning-curve analysis', *The Journal of Machine Learning Research* 4, pp. 211 - 255.

Petrovica, D., Royb, R. and Petrovicc, R. (1998) 'Modelling and simulation of a supply chain in an uncertain environment ', *European Journal of Operational Research,* 109, (2), pp. 299 - 309.

Regattieri, A., Gamberi, M., Gamberini, R. and Manzini, R. (2005) 'Managing lumpy demand for aircraft spare parts', *Journal of Air Transport Management,* 11, (6), pp. 426 - 431.

Rustenburg, W. D., Houtum, G. J. v. and Zijm, W. H. M. (2001) 'Spare parts management at complex technology-based organizations: an agenda for research', *International Journal of Production Economics,* 71, pp. 177 - 193.

Rustenburg, W. D., Van Houtum, G. J. and Zijm, W. H. M. (2000) 'Spare parts management for technical systems: resupply of spare parts under limited budgets', *IIE Transactions,* 32, (10), pp. 1013-1026.

Sani, B. and Kingsman, B. G. (1997) 'Selecting the best periodic inventory control and demand forecasting methods for low demand items', *Journal of the Operational Research Society,* 48, pp. 700 - 713.

Saunders, M., Lewis, P. and Thornhill, A. (2007) *Research Methods for Business Students, 4th edn.* Prentice Hall: Harlow, UK.

Saunders, S. (2009) *Jane's Fighting Ships 2009 - 2010.* Jane's Information Group: Coulsdon, Surrey, UK.

Schmitt, M. N. (2005) 'Precision attack and international humanitarian law', *International Review of the Red Cross,* 87, (839).

Schwarzkopf, A. B., Tersine, R. J. and Morris, J. S. (1988) 'Top-down versus Bottom-up Forecasting Strategies ', *International Journal of Production Research,* 26, (11), pp. 1833 - 1843.

Seon, M.-S. and U, J.-U. (2009) 'A study on forecasting of repair part demands of Korean Military: focused on Navy', *The Korean Journal of Defense Analysis,* 85, pp. 201 - 234.

Shlifer, E. and Wolff, R. W. (1979) 'Aggregation and proration in forecasting', *Management Science,* 25, (6), pp. 594 - 603.

Shumway, R. H. and Stoffer, D. S. (2006) *Time Series Analysis and Its Applications With R Examples, 2nd edn.* Springer: NY.

Silver, E. (1970) 'Some ideas related to the inventory control of items having erratic demand patterns', *Canadian Operational Research Journal,* 8, pp. 87 - 100.

Silver, E. A., Pyke, D. F. and Peterson, R. (1998) *Inventory Management and Production Planning and Scheduling, 3rd edn.* John Wiley & Son, Inc.: NY.

Sipila, J. (2004) *Terrorism and counter-terrorism: impact on defence and other security systems.* Helsinki: National Defence College

Slack, N., Chambers, S. and Johnson, R. (2004) *Operations Management, 4th edn.* Prentice Hall: Harlow, UK.

Sterman, J. D. (1989) 'Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment', *Management Science,* 35, (3), pp. 321 - 339.

Steyerberg, E. W., Harrell, F. E., Borsboom, G. J. J. M., Eijkemans, M. J. C. R., Vergouwe, Y. v. and Habbema, J. D. F. (2001) 'Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis', *Journal of Clinical Epidemiology,* 54, (8), pp. 774 - 781.

Syntetos, A. A. (2001) *Forecasting of Intermittent Demand.* Ph.D. Thesis. Brunel

University.

Syntetos, A. A. (2007) 'A note on managing lumpy demand for aircraft spare parts', *Journal of Air Transport Management*, 13, (3), pp. 166 - 167.

Syntetos, A. A. and Boylan, J. E. (2001) 'On the bias of intermittent demand estimates', *International Journal of Production Economics*, 71, (1 - 3), pp. 457 - 466.

Syntetos, A. A. and Boylan, J. E. (2005) 'The accuracy of intermittent demand estimates', *International Journal of Forecasting*, 21, (2), pp. 303 - 314.

Tabachnick, B. G. and Fidell, L. S. (2007) *Using Multivariate Statistics: 5th edn.* Pearson Education: London.

The US Naval Institute. (1977) *Naval Operations Analysis.* US Naval Institute: MD, USA.

Theil, H. (1954) *Linear Aggregation of Economic Relations.* North-Holland Publishing Company: Amsterdam.

Tu, J. V. (1996) 'Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes', *Journal of Clinical Epidemiology*, 49, (11), pp. 1225 - 1231.

Tukey, J. W. (1977) *Exploratory Data Analysis.* Addison-Wesley: Reading, MA.

Viswanathan, S., Widiarta, H. and Piplani, R. (2008) 'Forecasting aggregate time series with intermittent subaggregate components: top-down versus bottom-up forecasting', *IMA Journal of Management Mathematics*, pp. 1 - 13.

Voss, C., Tsikriktsis, N. and Frohlich, M. (2002) 'Case research in operations management', *International Journal of Operations & Production Management*, 22, (2), pp. 195 - 219.

Wacker, J. G. (1998) 'A definition of theory: research guidelines for different theory-building research methods in operations management', *Journal of Operations Management*, 16, (4), pp. 361 - 385.

Waller, D. L. (2003) *Operations Management: a Supply Chain Approach, 2nd edn.* Thompson: London.

Waters, C. D. J. (1991) *An Introduction to Operations Management.* Addison Wesley: Wokingham, UK.

Wei, W. W. S. and Abraham, B. (1981) 'Forecasting contemporal time series aggregates', *Communications in Statistics - Theory and Methods*, 10, (13), pp. 1335 - 1334.

Wemmerlöv, U. (1989) 'The behavior of lot-sizing procedures in the presence of forecast errors ', *Journal of Operations Management*, 8, (1), pp. 37 - 47

White, A. P. and Liu, W. Z. (1997) *Statistical Properties of Tree-based Approaches to Classification.* John Wiley & Sons: New York.

Widiarta, H., Viswanathan, S. and Piplani, R. (2006) 'On the effectiveness of top-down strategy for forecasting autoregressive demands', *Naval Research Logistics*, 54, (2), pp. 176 - 188.

Widiarta, H., Viswanathan, S. and Piplani, R. (2008a) 'Forecasting item-level demands: an analytical evaluation of top–down versus bottom–up forecasting in a production-planning framework ', *IMA Journal of Management Mathematics*, 19, pp. 207 - 218.

Widiarta, H., Viswanathan, S. and Piplani, R. (2009) 'Forecasting aggregate demand: an analytical evaluation of top-down versus bottom-up forecasting in a production planning framework', *International Journal of Production Economics*, 118, (1), pp. 87 - 94.

Widiarta, H., Viswanthan, S. and Piplani, R. (2008b) 'Evaluation of hierarchical forecasting for substitutable products', *International Journal of Services and Operations Management*, 4, (3), pp. 277 - 295.

Willemain, T. R., Smart, C. N. and Schwarz, H. F. (2004) 'A new approach to forecasting intermittent demand for service parts inventories', *International Journal of Forecasting*, 20, (3), pp. 375 - 387.

Willemain, T. R., Smart, C. N., Shockor, J. H. and DeSautels, P. A. (1994) 'Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method', *International Journal of Forecasting*, 10, pp. 529 - 538.

Williams, T. M. (1984) 'Stock control with sporadic and slow-moving demand', *Journal of Operational Research Society*, 35, (10), pp. 939 - 948.

Yin, R. K. (2003) *Case Study Research, 3rd edn.* Sage Publications: London, England.

Yonhap News (2007) *Military, the lack of maintenance budget.disabled fighting strength*. Available at: http://english.chosun.com/ (Accessed: 2 May 2007).

Yoon, K. B. and Sohn, S. Y. (2007) 'Forecasting both time varying MTBF of fighter aircraft module and expected demand of minor parts', *Journal of the Operational Research Society*, 58, (6), pp. 714 - 719.

Zotteri, G., Kalchschmidt, M. and Caniato, F. (2005) 'The impact of aggregation level on forecasting performance', *International Journal of Production Economics*, 93 - 94, (8), pp. 479 - 491.